On the random 2-stage minimum spanning tree

Abraham D. Flaxman
Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
email abie@cmu.edu

Alan Frieze *
Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
email alan@random.math.cmu.edu

Michael Krivelevich †
Department of Mathematics
Tel Aviv University
Tel Aviv 69978, Israel
email krivelev@post.tau.ac.il

December 20, 2004

Abstract

It is known [8] that if the edge costs of the complete graph K_n are independent random variables, uniformly distributed between 0 and 1, then the expected cost of the minimum spanning tree is asymptotically equal to $\zeta(3) = \sum_{i=1}^{\infty} i^{-3}$. Here we consider the following stochastic two-stage version of this optimization problem. There are two sets of edge costs $c_M \colon E \to \mathbb{R}$ and $c_T \colon E \to \mathbb{R}$, called Monday's prices and Tuesday's prices, respectively. For each edge e, both costs $c_M(e)$ and $c_T(e)$ are independent random variables, uniformly distributed in $[0, 1]$. The Monday costs are revealed first. The algorithm has to decide on Monday for each edge e whether to buy it at Monday's price $c_M(e)$, or to wait until its Tuesday price $c_T(e)$ appears. The set of edges X_M bought on Monday is then completed by the set of edges X_T bought on Tuesday to form a spanning tree. If both Monday's and Tuesday's prices were revealed simultaneously, then the optimal solution would have expected cost $\zeta(3)/2 + o(1)$. We show that in the case of two-stage optimization, the expected value of the optimal cost exceeds $\zeta(3)/2$ by an absolute constant $\epsilon > 0$. We also consider a threshold heuristic, where the algorithm buys on Monday only edges of cost less than α and completes them on Tuesday in an optimal way, and show that the optimal choice for α is $\alpha = 1/n$ with the expected cost $\zeta(3) - 1/2 + o(1)$. The threshold heuristic is shown to be sub-optimal. Finally we discuss the directed version of the problem, where the task is to construct a spanning out-arborescence rooted at a fixed vertex r, and show, somewhat surprisingly, that in this case a simple variant of the threshold heuristic gives the asymptotically optimal value $1 - 1/e + o(1)$.

*Supported in part by NSF Grant CCR-0200945.
†Supported in part by USA-Israel BSF Grant 2002-133, and by grant 64/01 from the Israel Science Foundation.

1 Introduction

Stochastic Programming refers to the general class of optimization problems where uncertainty is modelled by a probability distribution on the input variables. *Two stage optimization with recourse* is a widely used framework for stochastic optimization (see, e.g., the recent text by Birge and Louveaux [4]). In this paper we consider a particular example of this approach in the context of a basic combinatorial optimization problem.

The 2-stage spanning tree problem is defined as follows: We wish to find a low cost spanning tree of the complete graph K_n. On Monday, say, we are given edge costs, $c_M\colon E \to \mathbb{R}$. We also know that on Tuesday we will be given alternative costs for each edge, $c_T\colon E \to \mathbb{R}$. We do not know what the costs c_T will be, but they are random and we know their joint distribution $\pi(\omega)$, $\omega \in \Omega$, the set of possibilities. On Monday we must choose a set of edges X_M and pay for them at Monday's prices. On Tuesday, Monday's prices will no longer be available. Some edges will be cheaper and some will be more expensive. We must now choose a set of edges X_T, *at Tuesday's prices* to complete a spanning tree. Our total cost will be $c_M(X_M)+c_T(X_T)$, and our goal is to choose the set of edges X_M which minimizes the expected total cost of the tree we create. Formally, we wish to compute

$$OPT = \min_{X_M} \; c_M(X_M) + \mathrm{E}\left[\min_{X_T}\{c_T(X_T)\colon\; X_M \cup X_T \text{ is a spanning tree}\}\right].$$

Anupam Gupta [9] has pointed out that in the worst case, we can encode set cover as such a problem and so it probably cannot be efficiently approximated beyond a ratio of $O(\log n)$. A version of his proof is included in Appendix A.

The inapproximability result requires a worst-case set of costs c_M and a worst-case distribution for c_T. In this paper we carry out a probabilistic analysis for instances where $c_M(e)$ and $c_T(e)$, $e \in E(K_n)$ are selected independently and uniformly from the interval $[0, 1]$.

It is well-known that if only Monday's costs are available then we can find a minimum spanning tree in polynomial time and that the expected cost Z_1 of the optimum solution is asymptotically equal to $\zeta(3) \sim 1.20205\ldots$, Frieze [8]. Here $\zeta(3) = \sum_{n=1}^{\infty} n^{-3}$. Furthermore, if we could accurately predict the future and could find a minimum spanning tree using costs $c_2(e) = \min\{c_M(e), c_T(e)\}$ then [8] shows that we could pick edges so that our optimal cost Z_2 is asymptotically $\zeta(3)/2 \sim .601028\ldots$.

We first examine the performance of a simple threshold heuristic. Let $\mathcal{A}_\alpha$ be the algorithm that finds the minimum spanning forest of K_n that only uses eges of cost less than α on Monday and then completes the tree as cheaply as possible with new edges paid for at Tuesday's prices. Let A_α be the (random) value of the cost of the solution returned by $\mathcal{A}_\alpha$.

Theorem 1 *The best choice for α is $1/n$ in the sense that*

$$\mathrm{E}\left[A_{1/n}\right] = \zeta(3) - \frac{1}{2} + o(1) \leq \mathrm{E}\left[A_\alpha\right] + o(1)$$

for any choice of α.

Furthermore, for all α, the value A_α is concentrated around its mean.

The proof of Theorem 1 also gives a lower bound on the *value of stochastic solution (VSS)*, which is defined as the difference between the *expected result of using the expected value solution (EEV)* and the value of the optimal 2-stage solution OPT. To find the EEV, we observe that when the distribution of each Tuesday edge costs is replaced by its expected value, then the optimal solution **whp** ignores the Tuesday edges (which now have cost 0.5) and buys the whole tree on Monday. Thus, the EEV is asymptotically equal to the cost of buying the whole tree on Monday, which is asymptotically $\zeta(3)$. So we have $VSS = OPT - EEV \geq \frac{1}{2} - o(1)$.

Note that $\frac{\zeta(3)-1/2}{\zeta(3)/2} \sim 1.168\ldots$ and so **whp** [1] $A_{1/n}$ is within 17% of optimal.

Recently, Dhamdere and Singh showed that $\mathcal{A}_{\zeta(3)/n}$ is a constant factor approximation algorithm for instances where Monday's costs are arbitrary and Tuesday's costs are selected independently and uniformly between 0 and 1 [7].

A threshold algorithm is the best we can do if we do not take account of the structure of the costs for Monday's edges. Can we improve on this if we do? The answer is yes. We show that we can reduce the expected cost by at least a (very) small amount.

Theorem 2 *There is a polynomial time algorithm $\mathcal{A}^\star$ for selecting X_M whose (random) cost $A^\star$ satisfies*

$$\mathrm{E}\left[A^\star\right] \leq \zeta(3) - \frac{1}{2} - 10^{-256}.$$

We see therefore that the algorithm $\mathcal{A}_{1/n}$ is not optimal. Is it possible to asymptotically achieve $\zeta(3)/2$? Let OPT denote the minimum expected cost achievable by any 2-stage algorithm.

Theorem 3

$$OPT \geq \zeta(3)/2 + 10^{-5}$$

Theorem 3 is equivalent to a lower bound on the *expected value of perfect information (EVPI)*, which is defined between the difference between the value of the optimal 2-stage solution

[1] We use the term *with high probability*, abbreviated **whp**, to refer to a sequence of events $\{A_n\}$ for which $\Pr[A_n] \to 1$ as $n \to \infty$.

OPT and the expected value of the optimal solution when Tuesday's costs are known (the *wait-and-see value (WS))*. Theorem 3 shows that $EVPI = OPT - WS \geq 10^{-5}$.

Finding the the optimal choice of X_M and determining what can be done in polynomial time remain challenging open problems.

We continue with a directed version of this problem. Here we are given Monday and Tuesday costs for all the arcs of the complete digraph D_n. (The vertices of D_n are $\{1,\ldots,n\}$, and each ordered pair (i,j), $1 \leq i \neq j \leq n$, forms an arc in D_n). We now wish to find a low cost spanning arborescence rooted at vertex 1 i.e. a tree with arcs directed away from vertex 1. We first consider the threshold algorithm $\vec{\mathcal{A}}_\alpha$ which finds a minimum cost rooted forest using arcs from Monday of cost less than α and then completes it to a rooted arborescence after Tuesday's costs are revealed. Here $\alpha = 1/n$ is also the best choice: Let $\vec{A}_\alpha$ be the cost of the output from $\vec{\mathcal{A}}_\alpha$.

Theorem 4

$$\mathrm{E}\left[\vec{A}_{1/n}\right] = 1 - e^{-1} + o(1) \leq \mathrm{E}\left[\vec{A}_\alpha\right] + o(1)$$

for any choice of α.

Furthermore, for all α, the value $\vec{A}_\alpha$ is concentrated around its mean.

This turns out to be asymptotically optimal. Let $\overrightarrow{OPT}$ denote the minimum expected cost achievable by any 2-stage algorithm.

Theorem 5

$$\overrightarrow{OPT} \geq 1 - e^{-1} - o(1).$$

We prove Theorem 1 in Sections 2.1, 2.2, Theorem 2 in Section 2.3, Theorem 3 in Section 2.4, Theorem 4 in Section 3 and Theorem 5 in Section 3.2.

2 Undirected case

2.1 Threshold heuristic

Proof of Theorem 1 Fix some $0 < \alpha \leq 1$ and let T be the spanning tree produced by the threshold heuristic $\mathcal{A}_\alpha$, and let T_m and T_t be the edges bought on Monday and Tuesday respectively. Then

$$\begin{aligned}
c(T) &= \sum_{e \in T_m} c_m(e) + \sum_{e \in T_t} c_t(e) \\
&= \sum_{e \in T_m} \int_{p=0}^{1} \mathbf{1}_{\{c_m(e) \geq p\}} dp + \sum_{e \in T_t} \int_{p=0}^{1} \mathbf{1}_{\{c_t(e) \geq p\}} dp \\
&= \int_{p=0}^{1} \sum_{e \in T_m} \mathbf{1}_{\{c_m(e) \geq p\}} dp + \int_{p=0}^{1} \sum_{e \in T_t} \mathbf{1}_{\{c_t(e) \geq p\}} dp
\end{aligned}$$

For any graph G, let $\kappa(G)$ denote the number of connected components in G.

Now, let G_p be the graph containing only edges with Monday cost less than p. Since T_m is the minimum spanning forest on edges with Monday cost less than α, for $p \leq \alpha$ we have by the greedy algorithm of Kruskal:

$$\sum_{e \in T_m} \mathbf{1}_{\{c_m(e) \geq p\}} = \kappa(G_p) - \kappa(G_\alpha).$$

Let H_p be the graph containing edges with Monday cost less than α or Tuesday cost less than p. Then, since T_t is a minimum spanning tree on the graph formed by contracting each component of T_m to a single vertex, we have

$$\sum_{e \in T_t} \mathbf{1}_{\{c_t(e) \geq p\}} = \kappa(H_p) - 1.$$

Linearity of expectations gives

$$\begin{aligned}
\mathrm{E}\left[c(T)\right] &= \int_{p=0}^{\alpha} \mathrm{E}\left[\kappa(G_p) - \kappa(G_\alpha)\right] dp + \int_{p=0}^{1} \mathrm{E}\left[\kappa(H_p) - 1\right] dp \\
&= \int_{p=0}^{\alpha} \mathrm{E}\left[\kappa(G_p)\right] dp - \alpha\, \mathrm{E}\left[\kappa(G_\alpha)\right] + \int_{p=0}^{1} \mathrm{E}\left[\kappa(H_p)\right] dp - 1. \qquad (1)
\end{aligned}$$

We point out here that this implies

$$\int_{p=0}^{1} \mathrm{E}\left[\kappa(G_p)\right] dp = 1 + \zeta(3) + o(1), \qquad (2)$$

since putting $\alpha = 1$ into (1) we have $\mathrm{E}\left[c(T)\right]$ is the expected value of the minimum spanning tree using Monday costs. As already mentioned, this is $\zeta(3) + o(1)$. But we have $\kappa(G_\alpha) = \kappa(H_p) = 1$ for all $p \leq 1$.

Now, G_p is identically distributed with the Erdős-Rényi random graph $G_{n,p}$ (in which each pair of n vertices appears as an edge independently with probability p) and H_p is identically distributed with the Erdős-Rényi random graph $G_{n,p'}$ for $p' = \alpha + p - \alpha p$. So we have

$$\int_{p=0}^{1} \mathrm{E}\left[\kappa(H_p)\right] dp = (1-\alpha)^{-1} \int_{p'=\alpha}^{1} \mathrm{E}\left[\kappa(G_{p'})\right] dp'.$$

We may assume $\alpha \leq 2\log n/n$. If $\alpha > 2\log n/n$ then **whp** G_α is connected (see, for example, Bollobás [5, Thm. 7.3, p. 164]) which means that **whp** all the edges are purchased on Monday, and thus the expected cost $E[\mathcal{A}]_\alpha$ will be $\zeta(3) + o(1)$.

The integral $\int_{p'=\alpha}^{1} \mathrm{E}\left[\kappa(G_{p'})\right] dp'$ is bounded by $\int_{p'=0}^{1} \mathrm{E}\left[\kappa(G_{p'})\right] dp' = 1 + \zeta(3) + o(1)$, so we have (recalling that $\alpha = o(1)$)

$$\begin{aligned}\int_{p=0}^{\alpha} \mathrm{E}\left[\kappa(G_p)\right] dp + \int_{p=0}^{1} \mathrm{E}\left[\kappa(H_p)\right] dp &= \int_{p=0}^{1} E[\kappa(G_p)]dp - \frac{\alpha}{1-\alpha}\int_{p=\alpha}^{1} E[\kappa(G_p)]dp \\ &= \int_{p=0}^{1} \mathrm{E}\left[\kappa(G_p)\right] dp + o(1).\end{aligned}$$

So, altogether we have

$$\mathrm{E}\left[c(T)\right] = \zeta(3) + o(1) - \alpha\,\mathrm{E}\left[\kappa(G_\alpha)\right].$$

Set β so that $\alpha = \beta/n$ and put κ_T equal to the number of tree components. There are at most $n^{2/3}$ components of size at least $n^{1/3}$ and so we see that

$$\begin{aligned}\alpha\,\mathrm{E}\left[\kappa_T(G_\alpha)\right] &= \frac{\beta}{n}\sum_{k=1}^{n^{1/3}} \binom{n}{k} k^{k-2} \left(\frac{\beta}{n}\right)^{k-1} \left(1-\frac{\beta}{n}\right)^{k(n-k)+(k^2-3k+2)/2} + o(1) &\quad (3)\\ &= \sum_{k=1}^{\infty} \frac{k^{k-2}}{k!} \left(\beta e^{-\beta}\right)^k + o(1). &\quad (4)\end{aligned}$$

Note that the sum in (4) is convergent, even for $\beta = 1$.

Let κ_N denote the number of non-tree components. Then we have

$$\alpha\,\mathrm{E}\left[\kappa_N(G_\alpha)\right] \leq \frac{\beta}{n}\sum_{k=1}^{n^{1/3}} \binom{n}{k} k^{k} \left(\frac{\beta}{n}\right)^{k} \left(1-\frac{\beta}{n}\right)^{k(n-k)} + o(1) \leq \frac{\beta}{n}\sum_{k=1}^{n^{1/3}} (\beta e^{1-\beta})^k + o(1) = o(1),$$

and so

$$\alpha\,\mathrm{E}\left[\kappa(G_\alpha)\right] = \sum_{k=1}^{\infty} \frac{k^{k-2}}{k!} \left(\beta e^{-\beta}\right)^k + o(1). \quad (5)$$

Now $\beta e^{-\beta}$ has a unique maximum at $\beta = 1$, which shows that the threshold $\alpha = 1/n$ is asympotically best for the threshold heuristic.

Finally, we note that for $\beta = 1$,

$$\alpha\,\mathrm{E}\left[\kappa(G_\alpha)\right] = \sum_{k=1}^{\infty} \frac{k^{k-2}}{k!} e^{-k} + o(1) = \frac{1}{2} + o(1), \quad (6)$$

and so the threshold heuristic attains a value of $\zeta(3) - \frac{1}{2} + o(1)$.

(The last equation in (6) can be justified as follows: Consider the exponential generating function $U(x) = \sum_{k=1}^{\infty} \frac{k^{k-2}}{k!} x^k$ for the number of labelled trees with k vertices and the exponential generating function $T(x) = \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} x^k$ for the number of labelled rooted trees with k vertices. These satisfy $U(x) = T(x) - \frac{T(x)^2}{2}$ (equation (3.3) of [10]). Now $T(e^{-1}) = 1$ can be seen from the fact that $nT(e^{-1})$ is asymptotically equal to the number of vertices on trees in the random graph $G_{n,1/n}$. The sum in (6) is $U(e^{-1})$.)

2.2 Concentration

The goal of this section is to prove that for any constant $\lambda > 0$, there exists $\delta = \delta(\lambda) > 0$ such that for sufficiently large n,

$$\Pr[|A_\alpha - \mathrm{E}[A_\alpha]| \geq \lambda] \leq e^{-\delta n}.$$

We need only show this for λ sufficiently small, and it is convenient to define ε so that $\varepsilon + 4(1-\varepsilon)^{-1}(2\varepsilon + H(\varepsilon)) = \lambda$, where $H(x) = -x \ln x - (1-x)\ln(1-x)$ is the entropy function.

In our analysis we consider separately the contribution of long and short edges. Let $C = 2\varepsilon^{-1}$, and let Z denote the total cost of the edges used by $\mathcal{A}_\alpha$ with edge cost at most C/n. Let $N = 2\binom{n}{2}$ and note that Z is a function of N i.i.d. random variables $X_1, \ldots, X_N$ (one for each edge for each day). Also, each X_i is uniformly distributed on $[0,1]$.

We will show Z is concentrated using a variant of the Symmetric Logarithmic Sobolev Inequality from [6]. Let Z_i' denote the same quantity as Z, but with the variable X_i replaced by an independent copy X_i'. Then a simplified form of the Symmetric Logarithmic Sobolev Inequality [6, Corollary 3] says that if

$$\mathrm{E}\left[\sum_{i=1}^{N} (Z - Z_i')^2 \mathbf{1}_{Z > Z_i'} \;\middle|\; X_1, \ldots, X_N\right] \leq c$$

then for all $t > 0$,

$$\Pr[Z > \mathrm{E}[Z] + t] \leq e^{-t^2/4c},$$

and if

$$\mathrm{E}\left[\sum_{i=1}^{N} (Z_i' - Z)^2 \mathbf{1}_{Z_i' > Z} \;\middle|\; X_1, \ldots, X_N\right] \leq c$$

then for all $t > 0$,

$$\Pr[Z < \mathrm{E}[Z] - t] \leq e^{-t^2/4c}.$$

Changing the value of one edge can change the value of Z by at most C/n, so $(Z - Z_i')^2 < (C/n)^2$. Let I denote the indices of the edges which contribute to Z. If $i \notin I$ then $Z_i' < Z$

implies $X_i' \leq C/n$. So

$$\sum_{i=1}^{N}(Z - Z_i')^2 \mathbf{1}_{Z>Z_i'} \leq \sum_{i \in I}(C/n)^2 + \sum_{i \notin I}(C/n)^2 \mathbf{1}_{X_i'<C/n}.$$

Since there are less than n terms in the first sum and less than n^2 terms in the second sum, we have

$$\mathrm{E}\left[\sum_{i=1}^{N}(Z - Z_i')^2 \mathbf{1}_{Z>Z_i'} \;\middle|\; X_1, \ldots, X_N\right] \leq C^2/n + C^3/n \leq 2C^3/n.$$

If $i \notin I$ then we also have that $Z_i' > Z$ implies $X_i' \leq C/n$. So we also have

$$\mathrm{E}\left[\sum_{i=1}^{N}(Z_i' - Z)^2 \mathbf{1}_{Z_i'>Z} \;\middle|\; X_1, \ldots, X_N\right] \leq C^2/n + C^3/n \leq 2C^3/n.$$

Therefore,

$$\Pr\left[|Z - \mathrm{E}\left[Z\right]| \geq \varepsilon\right] \leq 2e^{-\varepsilon^2 n/8C^3} = 2e^{-\varepsilon^5 n/64}.$$

Let Z' denote the total cost of the edges used by $\mathcal{A}_\alpha$ with edge cost at least C/n. We will show that $Z' \geq \lambda - \varepsilon$ with exponentially small probability.

Let G be the graph containing edges with Monday or Tuesday cost less than C/n. Then G is identically distributed with $G_{n,p}$ for $p = 2C/n - (C/n)^2$. Let S denote the set of vertices that are not in the giant (more precisely, largest) component of G. We will obtain a exponential bound on the probability that $|S| \geq \varepsilon n$. To do so, we let $\mathcal{B}_1$ denote the event "there exists a set T such that $\varepsilon n \leq |T| \leq n/2$ and no edge of G crosses the cut between T and $\bar{T}$." Note that in order for $|S| \geq \varepsilon n$, it is necessary that event $\mathcal{B}_1$ holds: if $|\bar{S}| \geq \varepsilon n$, then (since $|S|$ also exceeds εn) either $T = S$ or $T = \bar{S}$ shows that $\mathcal{B}_1$ occurs; if $|\bar{S}| \leq \varepsilon n$, then all connected components of the graph have size at most εn and we can choose T to be the union of an appropriate collection of connected components.

Since $C = 2\varepsilon^{-1}$, we have

$$\Pr[|S| \geq \varepsilon n] \leq \Pr[\mathcal{B}_1] \leq \sum_{k=\varepsilon n}^{n/2} \binom{n}{k} \left(1 - \frac{C}{n}\right)^{2k(n-k)} \leq \sum_{k=\varepsilon n}^{n/2} e^{n-2Ck(1-k/n)} \leq ne^{-n}. \tag{7}$$

Z' can be bounded by the sum of (i) the edges of length $> C/n$ in the minimum spanning tree using Monday costs and (ii) the sum of the edges of length $> C/n$ in a minimum spanning tree of the graph obtained by shrinking the components of the Tuesday forest. (ii) is stochastically less than by (i). The sum in (i) can be bounded by the sum over the vertices $s \in S$ of the length of the cheapest edge from s to the giant component (more precisely largest component) of the graph spanned by the edges of length $< C/n$.

We finish by calculating an upper bound on the probability that any subset of size εn has the sum of the minimum cost edges exceeding $(\lambda - \varepsilon)/2$. Let V_1 denote the minimum of $n' := (1-\varepsilon)n$ independent random variables each uniformly distributed in $[0,1]$. Then $\mathrm{E}[V_1] = \frac{1}{n'+1}$, and

$$\mathrm{E}\left[e^{tV_1}\right] = \int_{x=0}^{1} e^{tx} n'(1-x)^{n'-1} dx = 1 + \sum_{k\geq 1} \frac{t^k}{n'(n'+1)\cdots(n'+k-1)} \leq \left(1 + \frac{2t}{n'}\right).$$

(The second equality follows from integration by parts, inequality holds for $t \leq n'/2$).

Then, for any set T with $|T| = k$,

$$\Pr\left[\sum_{v\in T} V_1(v) \geq \lambda\right] = \Pr\left[e^{\frac{n'}{2}\sum_{v\in T} V_1(v)} \geq e^{\lambda n'/2}\right] \leq e^{-\lambda n'/2}\,\mathrm{E}\left[e^{n' V_1/2}\right]^k \leq e^{-\lambda n'/2 + k}.$$

Let $\mathcal{B}_2$ denote the event "there exists a set T with $|T| \leq \varepsilon n$ and $\sum_{v\in T} V_1(v) \geq (\lambda - \varepsilon)/2 = 2(1-\varepsilon)^{-1}(2\varepsilon + H(\varepsilon))$". Then we have

$$\Pr[\mathcal{B}_2] \leq \sum_{1\leq k\leq \varepsilon n} \binom{n}{k} e^{-\varepsilon n - H(\varepsilon)n} \leq \varepsilon n e^{-\varepsilon n}. \tag{8}$$

We combine (7) and (8) to show that the probability Z' exceeds $\lambda - \varepsilon$ is small.

$$\Pr\left[Z' \geq \lambda - \varepsilon\right] \leq \Pr[|S| \geq \varepsilon n] + 2\Pr[\mathcal{B}_2] \leq ne^{-n} + 2\varepsilon n e^{-\varepsilon n}.$$

Finally,

$$\begin{aligned}\Pr[|\mathcal{A}_\alpha - \mathrm{E}[\mathcal{A}_\alpha]| \geq \lambda] &\leq \Pr[|Z - \mathrm{E}[Z]| \geq \varepsilon] + \Pr[Z' \geq \lambda - \varepsilon] \\ &\leq 2e^{-\varepsilon^5 n/64} + ne^{-n} + 2\varepsilon n e^{-\varepsilon n}.\end{aligned}$$

2.3 Beyond the threshold heuristic

We can achieve a slightly better expected value than the threshold heuristic $\mathcal{A}_{1/n}$ by being more careful about edges with cost near the threshold.

Let ℓ be a positive integer and let $\varepsilon > 0$ be a small positive constant and let F be the minimum spanning forest on the edges with Monday cost less than $(1-\varepsilon)/n$. Let an edge $e = \{u, v\}$ be *bad* if it has Monday cost $c_M(e) \in [(1-\varepsilon)/n, 1/n]$, and for $x = u, v$ there are:

(A) Exactly ℓ vertices w for which $c_M(x, w) < (1-2\varepsilon)/n$. Denote this set of vertices by C_x.

(B) No vertices w for which $c_M(x, w) \in [(1-2\varepsilon)/n, 1/n]$.

(C) No vertices $w \in C_x$ and $y \notin \{x\} \cup C_x$ for which $c_M(y, w) < 1/n$.

If e is bad then e will be part of an isolated tree of $G_{1/n}$ containing $2\ell + 1$ edges and e will be chosen by $\mathcal{A}_{1/n}$.

Let T_1 be the tree constructed by $\mathcal{A}_{1/n}$ and let T_2 be obtained by taking the minimum spanning forest which uses edges e with $c_M(e) < 1/n$ which are not bad, and then completing this forest to a tree as cheaply as possible with edges at Tuesday's costs. We will show that

$$\mathrm{E}\left[T_1 - T_2\right] \geq 10^{-256} \tag{9}$$

and so completing the proof of Theorem 2.

We must estimate the expected savings if we leave out the bad edges and only the bad edges from the threshold solution. In this case, $\{x\} \cup C_x$, $x = u, v$ are trees of the forest of the edges chosen on Monday.

We consider the contribution from the removal of a single bad edge $e = \{u, v\}$. We expose the costs of the edges carefully to avoid unpleasant conditioning. First we expose the Monday cost of e. The probability $c_M(e)$ is in the correct range is ε/n. If $c_M(e)$ is in this range, we expose the Monday costs of the other edges incident to u and v. The probability that the costs of the other edges are in the correct range is

$$\left(\binom{n-1}{\ell}\left(\frac{1-2\varepsilon}{n}\right)^{\ell}\left(1-\frac{1}{n}\right)^{n-2-\ell}\right)^2 \geq \frac{(1-2\varepsilon)^{2\ell}}{e^2(\ell!)^2}(1-o(1)).$$

Now, we expose the Monday costs of the neighbors of $C_u \cup C_v$. The probability that (C) holds is $(1 - 1/n)^{2\ell(n-2-2\ell)} = e^{-2\ell}(1 + o(1))$.

Thus the expectation of the number of bad edges b is given by

$$\mathrm{E}\left[b\right] = (1 + o(1))\frac{\varepsilon(1-2\varepsilon)^{2\ell}e^{-2\ell-2}}{2(\ell!)^2}n. \tag{10}$$

We now expose all the Monday and Tuesday costs between the $n - 2 - 2\ell$ vertices that are not part of C_u and C_v. Let H be the graph containing all edges just exposed with Monday or Tuesday cost at most $(1 - 2\varepsilon)/n$. Note that H is identically distributed with $G_{n',p}$ for $n' = n - 2 - 2\ell$ and $p = (1 + o(1))(2 - 4\varepsilon)/n$. If $\varepsilon < 1/4$ then H has a giant component K_H **qs**[2]. We expose the remaining edge costs and let X_u (resp. X_v) be the minimum cost of a Tuesday edge from C_u (C_v) to K_H, assuming that it exists. The size of K_H is at least $\beta n(1-o(1))$ **qs**, where β is the root of $\beta + e^{-2(1-2\varepsilon)\beta} = 1$ in the interval $(0, 1)$. We take $\varepsilon = 0.1$ and then $\beta > 0.7$. So, for $\ell = 100$ we have $E[X_u] = E[X_v] = \frac{1+o(1)}{\ell\beta n} \leq 0.02n^{-1}$. For each bad edge $e = \{u, v\}$ we then have expected cost savings of at least

$$\frac{1-\varepsilon}{n} - \max\left\{\frac{1-2\varepsilon}{n}, X_u + X_v\right\} \geq \frac{0.1}{n}. \tag{11}$$

[2] A sequence of events $\mathcal{E}_n$ occurs *quite surely* **qs** if $\Pr(\mathcal{E}_n) = 1 - O(n^{-K})$ for any $K > 0$.

We can prove (11) as follows: Let $e = \{u, v\}$ be bad. $e \notin T_2$ and there is a path from u to v which goes to a vertex of C_u, goes to H via an edge of length X_u, traverses H and then goes via an edge of length X_v to a vertex of C_v and then to v. If A, B are the components of $T_1 - e$ then at least one edge $f \notin T_1$ of P will join A to B. We observe that

$$\min\{c_M(f), c_T(f)\} \leq \max\left\{\frac{1-2\varepsilon}{n}, X_u, X_v\right\} \leq \max\left\{\frac{1-2\varepsilon}{n}, X_u + X_v\right\}.$$

So, if we replace e by f in T_1 we will, by (11), save at least $\frac{0.1}{n}$. If we repeat this for all bad edges, then we will have a tree containing all of the Monday purchased edges and it will, in expectation, be at least $\frac{0.1\,\mathrm{E}[b]}{n}$ cheaper. We obtain (9) by using this together with (10) with $\ell = 100$.

2.4 A lower bound on OPT

If we could see all the Monday and Tuesday costs before selecting any edge then we could find a spanning tree with cost $\sim \zeta(3)/2$. Since we have to make some decisions before we see the Tuesday costs, it seems likely that our solution should, in expectation, cost at least $\zeta(3)/2 + \varepsilon$, for some small ε. This is the content of Theorem 3.

Let C be a positive constant, (which we will eventually take to be 3, to obtain a concrete bound). Consider the edges we buy on Monday with cost exceeding $\frac{C}{n}$. Let

$$\varepsilon = \beta_C e^{-(2C+3)}/2$$

where β_C is the solution to $\beta + e^{-(C-1)\beta} = 1$ in the interval $(0, 1)$.

We will see that if we buy more than εn of these edges, then we will regret our purchase on Tuesday. We also argue that if we buy less than εn, then we will regret it too.

Case 1: Suppose X_M contains at least εn edges with $c_M(e) \geq \frac{C}{n}$, and let $e_1, e_2, \ldots, e_m$, $m \geq \varepsilon n$ be these edges. Let H be the graph consisting of all the edges e' with $c_T(e') < \frac{C-1}{n}$. Then (for any $C > 2$), H contains a giant component K_H with size $\beta_C n(1 - o(1)$ **whp**. For $i = 1, \ldots, m$, if e_i has both end vertices in K_H, then we can find a cheaper spanning tree T_i by removing e_i from T_{i-1} and adding an edge from H on Tuesday. This will decrease the cost of the solution by at least $1/n$. Since each edge e_i has both vertices in K_H with probability $\sim \binom{\beta_C n}{2}/\binom{n}{2} \sim \beta_C^2$, the 2-stage solution exceeds the optimal solution by at least $\beta_C^2 \varepsilon - o(1)$ in expectation.

Case 2: Suppose X_M contains less than εn edges with $c_M(e) \geq \frac{C}{n}$. For a vertex v, let $\mathcal{E}_v$ be the event "the cheapest Monday edge incident to v has cost between $\frac{C}{n}$ and $\frac{C+1}{n}$ and the other endpoint is in K_H". Then $\Pr\left[\mathcal{E}_v \mid |K_H|\right] = |K_H| \frac{1}{n}\left(1 - \frac{C+1}{n}\right)^{n-2}$, and so $\Pr[\mathcal{E}_v] \sim \beta_C e^{-(C+1)}$.

Let $\mathcal{E}'_v$ be the event "there is no edge incident to v with Tuesday cost less than $\frac{C+2}{n}$". Then $\Pr[\mathcal{E}'_v] = \left(1 - \frac{C+2}{n}\right)^{n-1} \sim e^{-(C+2)}$. If $\mathcal{E}_v$ and $\mathcal{E}'_v$ occur then not buying the edge from v to K_H

with cost less than $(C+1)/n$ on Monday results in paying at least $\frac{1}{n}$ more than optimal to connect v on Tuesday. But we only take εn edges on Monday with $c_M(e) \geq \frac{C}{n}$, so we expect to pay this penalty on at least $n\beta_C e^{-(2C+3)} - \varepsilon n$ vertices, and so our 2-stage solution exceeds optimal by at least $\beta_C e^{-(2C+3)}/4$ in expectation, after accounting for the fact that one edge has 2 endpoints.

Taking $C = 3$, numerical computation shows that the 2-stage solution exceeds optimal by at least 10^{-5}.

3 Spanning arborescence problem

The directed version of this problem is to build a cheap spanning out-arborescence rooted at a fixed vertex r. Given a random cost for each directed edge on Monday and a distribution for the random cost for each directed edge on Tuesday, find directed edges to buy on Monday to minimize the expected total cost when you buy the missing edges on Tuesday. In other words, compute

$$OPT = \min_{X_M} c_M(X_M) + \mathrm{E}\left[\min_{X_T}\{c_T(X_T) : X_M \cup X_T \text{ is a spanning arborescence rooted at } r\}\right].$$

In this case there is a lower bound that matches the threshold heuristic.

3.1 Threshold heuristic

This comprises two phases:

Phase 1: For each vertex, if the cheapest in-edge on Monday has cost at most α we will buy it, and otherwise we will wait till Tuesday and buy the cheapest in-edge available. This does not define an arborescence, it defines a functional digraph, with all in-degrees equal to 1. This consists of a collection of vertex disjoint cycles $C_1, C_2, \dots, C_m$, and for each vertex v in $C_1 \cup C_2 \cup \cdots \cup C_m$ there is an arborescence directed from v.

Phase 2: We delete the arc directed into r. We then delete one (arbitrary) arc from each cycle that remains. At this point we have m' vertex disjoint directed rooted trees $T_1, T_2, \dots, T_{m'}$, say, where $m \leq m' \leq m+1$. Assume that r is the root of T_1. Now we make a spanning arborescence as follows:
For $i = m', m'-1, \dots, 2$ we do the following:
Find the cheapest arc, at Tuesday's prices, into T_i from a vertex in $T_1 \cap T_2 \cap \ldots \cap T_{i-1}$.
If this arc came from T_j then this creates a rooted tree T_j' from the vertices of T_j, T_i.
T_j' replaces T_j and T_i disappears.

Note that since the arc removed in Phase 2 is chosen arbitrarily, this procedure can be implemented in the 2-Stage framework: on Monday, we leave some edge out of any cycle that Phase 1 wants to buy. This does not require any knowledge of the Tuesday costs.

Analysis of Phase 1
We find that the expected cost of the arcs chosen is given by

$$n\left(\int_{x=0}^{\alpha}(n-1)x(1-x)^{n-2}dx+(1-\alpha)^{n-1}\int_{x=0}^{1}(n-1)x(1-x)^{n-2}dx\right)$$
$$=n\left(\frac{1}{n}-\frac{(1-\alpha)^{n-1}}{n}(1+\alpha n-\alpha)+\frac{(1-\alpha)^{n-1}}{n}\right)=1-(n-1)\alpha(1-\alpha)^{n-1}.$$

This is minimised at $\alpha=1/n$ giving a value which is asymptotically equal to $1-e^{-1}$.

Analysis of Phase 2
It remains to show that the cost added in this phase is $o(1)$ **whp**. First of all, it is known (see, e.g., [5], Ch. 14.5), that for some $K>0$, $m\le K\log n$ with probability at least $1-O(n^{-2})$. An easy calculation shows that with probability $1-o(n^{-2})$ over Tuesdays' prices, for every ordered partition (V_1,V_2) of V the cheapest Tuesday's arc from V_1 to V_2 has cost at most $\frac{4\log n}{n}$. Indeed, the probability that this is not so can be bounded from above by

$$\begin{aligned}\sum_{k=1}^{n-1}\binom{n}{k}\left(1-\frac{4\log n}{n}\right)^{k(n-k)} &\le 2\sum_{k=1}^{n/2}\binom{n}{k}\left(1-\frac{4\log n}{n}\right)^{k(n-k)}\le 2\sum_{k=1}^{n/2}\left(\frac{en}{k}\right)^{k}e^{-\frac{4\log n}{n}\cdot\frac{kn}{2}}\\ &= o(n^{-2})\ .\end{aligned}$$

Assuming the above conditions hold the arcs added at Phase 2 increase the total weight of the obtained solution by at most $O(\frac{\log^2 n}{n})$.

Concentration
The proof is analogous to the proof in section 2.2. Given a $\lambda>0$, we pick the appropriate constant C, and use Azuma's inequality to show the total cost of the arcs with cost less than C/n is concentrated around its mean. Then we show that the probability the total cost of the remaining arcs is anything significant is exponentially small, by showing that (with probability exponentially close to 1) there are not too many vertices left unconnected, and for any small set of vertices, there is a set of edges connecting them to the remaining vertices which doesn't cost anything significant.

3.2 Matching lower bound on $\overrightarrow{OPT}$

In any feasible solution each vertex v besides the root r has to have a unique edge directed to it. So we can obtain a lower bound on what is achievable by looking at each vertex individually.

For any realization of c_M, taking expectations over c_T we have

$$\min_{X_M}\left\{c_M(X_M)+\mathrm{E}\left[\min_{X_T} c_T(X_T):\ X_M\cup X_T \text{ is an arborescence}\right]\right\}$$
$$\geq \sum_{v\neq r}\min\left\{\min_{w\neq v}\{c_M(w,v)\},\mathrm{E}\left[\min_{w\neq v}\{c_T(w,v)\}\right]\right\}=\sum_{v\neq r}\min\left\{\min_{w\neq v}\{c_M(w,v)\},1/n\right\}.$$

And so taking expectations over c_M, we obtain, where X_i are independent, uniformly distributed between 0 and 1,

$$\begin{aligned}\overrightarrow{OPT} &\geq (n-1)\,\mathrm{E}\left[\min\{1/n,X_1,X_2,\ldots,X_{n-1}\}\right]\\ &= (n-1)\left(\int_{x=0}^{1/n}(n-1)x(1-x)^{n-2}dx+\left(1-\frac{1}{n}\right)^{n-1}\int_{x=1/n}^{1}\frac{1}{n}dx\right)\\ &\sim 1-e^{-1}.\end{aligned}$$

4 Open questions

As far as this piece of work is concerned, the main open question is how to close the gap between the results of Theorems 2 and 3.

Another natural question might be to consider a 2-stage version of the random assignment problem. See Aldous [1], [2], Linusson and Wästlund [12] and Nair, Prabhakar and Sharma [13] for recent work on the standard *one-stage* analysis. In principal, one could try to carry out a similar 2-stage probabilistic analysis for any combinatorial optimization problem.

References

[1] D. Aldous, Asymptotics in the random assignment problem, *Probability Theory and Related Fields* 93 (1992) 507-534.

[2] D. Aldous, The $\zeta(2)$ limit in the random assignment problem, *Random Structures and Algorithms* 18 (2001) 381-418.

[3] A. Beveridge, A.M. Frieze and C.J.H. McDiarmid, Random minimum length spanning trees in regular graphs, *Combinatorica* 18 (1998) 311-333.

[4] J. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer, 1997.

[5] B.Bollobás, Random Graphs, (2nd Edition) Cambridge University Press (2001).

[6] S. Boucheron, G. Lugosi, and P. Massart, Concentration inequalities using the entropy method, *Annals of Probability*, 31 (2003) 1583-1614.

[7] K. Dhamdere and M. Singh, Handling the random tree, manuscript.

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

[9] A. Gupta, private communication.

[10] S.Janson, T.Łuczak, D.E. Knuth and B.G. Pittel, The birth of the giant component, *Random Structures and Algorithms* 4 (1993) 233-357.

[11] S.Janson, T.Łuczak and A.Ruciński, *Random Graphs*, John Wiley and Sons, (2000).

[12] S. Linusson and J. Wästlund, A solution of Parisi's conjecture on the random assignment problem, *Probability Theory and Related Fields* 128 (2004), 419–440.

[13] C. Nair, B. Prabhakar and M. Sharma, Proofs of the Parisi and Coppersmith-Sorkin Conjectures for the Finite Random Assignment Problem, *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computing* (2003), 168–177.

A Hardness of approximation in worst case

We describe a gap preserving reduction from set cover. Let $S_1, S_2, \ldots, S_m \subseteq [n]$ be a set cover instance. We construct an MST instance with $n+m+1$ vertices by defining the function c_M and the random function c_T. Denote the vertices by $\{r, v_1, \ldots, v_m, 1, \ldots, n\}$. Set the Monday edge cost of $\{r, v_i\}$ to 1 and set all the other Monday edge costs to ∞.

$$c_M(\{u,v\}) = \begin{cases} 1, & \text{if } \{u,v\} = \{r, v_i\}; \\ \infty, & \text{otherwise.} \end{cases}$$

Make the Tuesday edge costs uniformly distributed over n functions, where the j-th function sets to ∞ the cost of edges in the cut separating $T_j = \{j\} \cup \{v_i\colon\ S_i \ni j\}$ from the rest of the graph, and sets the other edges costs to 0.

$$c_T^{(j)}(\{u,v\}) = \begin{cases} \infty, & \text{if } \{u,v\} \in (T_j, \overline{T_j}); \\ 0, & \text{otherwise.} \end{cases}$$

If $S_{i_1} \cup S_{i_2} \cup \cdots \cup S_{i_k} = [n]$ then by buying Monday edges $\{r, v_{i_j}\}$ where $j = 1, \ldots, k$, we can complete the spanning tree on Tuesday with 0-cost edges for any future.

On the other hand, consider any set X_M of Monday edges such that the expected total cost of the spanning tree is finite. Then each $\{u,v\} \in X_M$ must have the form $\{r, v_{i_j}\}$. Consider set of sets corresponding to these edges, $\{S_{i_1}, \ldots, S_{i_k}\}$. For any $\ell \in [n]$, we must have $\ell \in S_{i_j}$ for some i_j; otherwise with probability $1/n$, we realize future ℓ, and have to buy an infinite cost edge across cut $(T_\ell, \overline{T_\ell})$.