Available online at www.sciencedirect.com
ScienceDirect
Theoretical Computer Science 384 (2007) 161–167
Theoretical Computer Science
www.elsevier.com/locate/tcs

On the bounded-hop MST problem on random Euclidean instances[☆]

Andrea E.F. Clementi[a,*], Miriam Di Ianni[a], Massimo Lauria[b], Angelo Monti[b], Gianluca Rossi[a], Riccardo Silvestri[b]

[a] *Dipartimento di Matematica, Università degli Studi di Roma "Tor Vergata", Italy*
[b] *Dipartimento di Informatica, Università degli Studi di Roma "La Sapienza", Italy*

Abstract

The d-DIM h-HOPS MST problem is defined as follows: given a set S of points in the d-dimensional Euclidean space and $s \in S$, find a minimum-cost spanning tree for S rooted at s with height at most h. We investigate the problem for any constant h and $d > 0$. We prove the first nontrivial lower bound on the solution cost for almost all Euclidean instances (i.e. the lower bound holds with high probability). Then we introduce an easy-to-implement, fast divide et impera heuristic and we prove that its solution cost matches the lower bound.

Keywords: Approximation algorithms; Randomized algorithms; Bounded height minimum spanning tree

1. Introduction

Given a positive integer h, an *h-tree* T is a rooted tree such that the number of hops (edges) in the path from the root to any other node is not greater than h. The *cost* of T, denoted as $\mathsf{cost}(T)$, is the sum of its edge weights. The *Minimum h-hops Spanning Tree* problem (h-HOPS MST) is defined as follows: Given a graph $G(V, E)$ with nonnegative edge weights and a node $s \in V$, find a minimum-cost h-tree rooted at s and spanning V. The h-HOPS MST problem and the related problem in which the constraint is on the tree diameter find applications in several areas: Networks [4], distributed system design [21,7] and bit-compression for information retrieval [6].

The efficient construction of a (minimum) spanning tree of a communication network yields good protocols for *broadcast* and *antibroadcast*[1] operations. The hop restriction limits the maximum number of links or connections in the communication paths between source and destination nodes: It is thus closely related to restricting the maximum delay transmission time of such fundamental communication protocols. The hop restriction finds another relevant application in the context of reliability: Assume that, in a communication network, link faults happen with probability p and that all faults occur independently. Then, the probability that a multi-hop transmission fails exponentially

☆ Supported by the European Union under the Integrated Project IST-15964 AEOLUS ("Algorithmic Principles for Building Efficient Overlay Computers").

* Corresponding author. Tel.: +39 06 72594622; fax: +39 06 72594699.
E-mail addresses: clementi@mat.uniroma2.it (A.E.F. Clementi), diianni@mat.uniroma2.it (M. Di Ianni), lauria@di.uniroma1.it (M. Lauria), monti@di.uniroma1.it (A. Monti), gianluca.rossi@uniroma2.it (G. Rossi), silver@di.uniroma1.it (R. Silvestri).

1 The antibroadcast operation is also known in literature as *Accumulation* or *All-to-One* operation.

0304-3975/$ - see front matter
doi:10.1016/j.tcs.2007.04.039

increases with the number of hops. Summarizing, a fixed bound on the maximum number of hops is sometimes a necessary constraint in order to achieve fast and reliable communication protocols. For further motivations in studying the h-HOPS MST problem see [5,11,15,23].

In [1] Alfandari and Paschos proved that METRIC 2-HOPS MST (i.e. the problem version where edge weights of the input graph yield a metric) is NP-hard and no PTAS exists unless P = NP. The first constant factor approximation algorithm was given by Shmoys et al. in [22]: they presented a 3.16 approximation algorithm. After this, a series of constant factor approximation algorithms was published, see [8,17,14]. Currently, the best factor is 1.52 due to Mahdian et al. [18]. All such algorithms are not practically efficient.

Several previous works [5,11,23] focused on the h-HOPS MST problem version (and some generalizations) where nodes are points of the Euclidean 2-dimensional space, the graph is complete, and the edge weights are the Euclidean distances. This problem version will be called 2-DIM h-HOPS MST. As for the case $h = 2$, the problem can be easily reduced to the classic Facility Location Problem on the plane. Indeed, the distance of the root from vertex i can be seen as the cost of opening a facility at vertex i. It thus follows that all the approximation algorithms for the latter problem apply to the 2-DIM 2-HOPS MST as well. In particular, the best result is the PTAS given by Arora et al in [3]. The algorithm works also in higher dimensions; however, it is based on a complex dynamic programming technique that makes any implementation very far to be practical. For $h \geq 2$, neither hardness results nor polynomial-time (exact) algorithms are known for the 2-DIM h-HOPS MST problem. Even more, for $h \geq 3$, no polynomial-time, constant-factor approximation algorithms are known.

Another series of papers have been devoted to evaluating and comparing solutions for the 2-DIM h-HOPS MST problem returned by some heuristics on random 2-dimensional instances by performing computer experiments [9,20, 23]. Almost all such works adopt the *uniform input random model*, i.e. points are chosen independently and uniformly at random from a fixed square of the plane. The motivation on this input model is twofold: on the one hand, the uniform distribution is the most suitable choice when nothing is known about the real input distribution or when the goal is to perform a preliminary study of the heuristic on arbitrary instances. On the other hand, uniform distribution well models important applications in the area of ad-hoc wireless and sensor networks. In such scenarios, once base stations are efficiently located, a large set of small wireless (mobile or not) devices are *well-spread* over a geographical region. Needless to say, efficient and reliable protocols for broadcast and accumulation are a primary goal [10] for such networks. We emphasize that no theoretical analysis is currently available on the expected performance of any efficient algorithm for the 2-DIM h-HOPS MST problem.

In [2,16] a polynomial-time $O(\log n)$-approximation algorithm is given for the h-HOPS MST problem, but its time complexity is $n^{O(h)}$. Gouveia in [11,12] and Gouveia and Requejo [13] provided lower bounds on the optimal cost of the h-HOPS MST based on integer programming models. Voss in [23] presented a tabu-search heuristic for the h-HOPS MST problem but its time complexity is very high when the graph is dense. In [20] heuristics based on Prim's algorithm and Evolutionary techniques have been experimentally tested. Finally, in [9] experimental tests have been performed on greedy heuristics and on the one analysed in this paper.

In the sequel, with the term *random set of points*, we mean a finite set of points chosen independently and uniformly at random from a fixed d-dimensional hypercube (d-cube).

Our first result is a lower bound on the cost of any h-tree spanning a random set of points.

Theorem 1. *Let $h, d \geq 1$ be constants. Let S be a random set of n points in a d-cube of side length L. Then, with high probability, for any h-tree T spanning S, it holds that*

$$\mathsf{cost}(T) = \begin{cases} \Omega\left(L \cdot n^{\frac{1}{h}}\right) & \text{if } d = 1 \\ \Omega\left(L \cdot n^{1-\frac{1}{d}+\frac{d-1}{d^{h+1}-d}}\right) & \text{otherwise.} \end{cases}$$

Here and in the sequel the term *with high probability* (in short, *w.h.p.*) means that the event holds with probability at least $1 - e^{-c \cdot n}$, for some constant $c > 0$. So, according to our input model, claiming that a given bound holds w.h.p. is equivalent to claiming that it holds *for almost all* inputs.

We then introduce a simple Divide et Impera heuristic denoted as h-PARTY. It makes a partition into cells of the smallest d-cube containing S. In each nonempty cell, it selects an arbitrary subroot s' and connects s' to the root s; finally, it recursively solves the nonempty cell subinstances of the problem with $h - 1$ hops. Choosing the "correct" size of the cells is the critical technical issue. This is solved thanks to the lower bound function in Theorem 1.

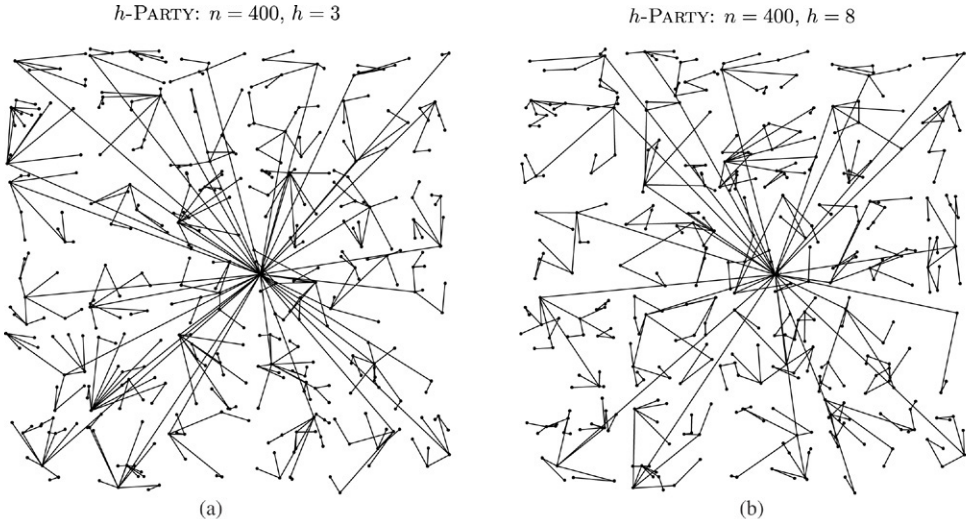

Fig. 1. The spanning trees yielded by the h-PARTY heuristics on the same random set of 400 points and $h = 3, 8$.

Theorem 2. *Let $h, d \geq 1$ be constants. Let S be a set of n points in a d-cube of side length L and let $s \in S$. For any h-tree T returned by h-PARTY on input (S, s), it holds that*

$$\mathsf{cost}(T) = \begin{cases} \mathrm{O}\left(L \cdot n^{\frac{1}{h}}\right) & \text{if } d = 1 \\ \mathrm{O}\left(L \cdot n^{1-\frac{1}{d}+\frac{d-1}{d^{h+1}-d}}\right) & \text{otherwise.} \end{cases}$$

Theorems 1 and 2 imply that, for any fixed h, h-PARTY returns a solution which is, with high probability, a constant factor approximation of the optimum. So, even though this fast algorithm provides no provably-good approximation in the worst case, it works well on almost all Euclidean instances.

h-PARTY is the first heuristic for the 2-DIM h-HOPS MST that works in $O(n)$ time and it can be thus efficiently applied to very large instances. In fact, the heuristic has been implemented and tested on instances of hundreds of thousands of points [9] (see Fig. 1 for two outputs of the heuristic).

Notice that, differently from Theorem 1, the bound in Theorem 2 holds for *any* Euclidean instance. It thus follows that random instances are those having the largest cost.

2. Preliminaries

In the proof of our results we make use of the well-known Hölder inequality. We present it in the following convenient forms. Let x_i, $i = 1, \ldots, k$ be a set of k nonnegative reals and let $p, q \in \mathcal{R}$ such that $p \geq 1$ and $q \leq 1$. Then, it holds that

$$\sum_{i=1}^{k} x_i^p \geq k \left(\frac{\sum_{i=1}^{k} x_i}{k} \right)^p ; \tag{1}$$

$$\sum_{i=1}^{k} x_i^q \leq k \left(\frac{\sum_{i=1}^{k} x_i}{k} \right)^q . \tag{2}$$

3. The lower bound

Next lemma provides the first known deterministic lower bound on the cost of h-trees for general Euclidean instances.

Lemma 3. *Let $h, d \geq 1$ be constants. Let S be a set of points in a d-dimensional Euclidean space. Consider a partition of the space into d-cubes, with the side length of each d-cube being l, and let n_l be the number of the d-cubes containing some point of S. For any h-tree T spanning S it holds that*

$$\mathsf{cost}(T) = \begin{cases} \Omega\left(l \cdot n_l^{1+\frac{1}{h}}\right) & \text{if } d = 1 \\ \Omega\left(l \cdot n_l^{1+\frac{d-1}{d^{h+1}-d}}\right) & \text{otherwise.} \end{cases}$$

Proof. Let

$$g(h) = \begin{cases} d & \text{if } h = 1 \\ d \cdot g(h-1) + d & \text{otherwise} \end{cases}$$

then,

$$\frac{1}{g(h)} = \begin{cases} \frac{d-1}{d^{h+1}-d} & \text{if } d > 1 \\ \frac{1}{h} & \text{if } d = 1. \end{cases}$$

Hence, we aim to show that $\mathsf{cost}(T) = \Omega\left(l \cdot n_l^{1+\frac{1}{g(h)}}\right)$.

Let s be the root point of the spanning tree T and consider a d-sphere centred at s and of radius $r = \Theta(l \cdot (n_l)^{\frac{1}{d}})$ such that the number n_l' of nonempty d-cubes outside the sphere is at least $n_l/2$. Finally let B be the set of points in these n_l' d-cubes.

The proof is by induction on the height h of the spanning tree T. If $h = 1$, for each of the n_l' d-cubes, there is an edge in T of length at least r. This implies that

$$\mathsf{cost}(T) \geq r \cdot n_l' = \Omega\left(l \cdot n_l^{1+\frac{1}{d}}\right) = \Omega\left(l \cdot n_l^{1+\frac{1}{g(1)}}\right).$$

Let $h \geq 2$. Let $A = \{a_1, a_2, \ldots, a_{|A|}\}$ be the set of points whose father in T is at distance at least $\frac{r}{h}$ and let

$$\beta = 1 - \frac{1}{d} + \frac{1}{g(h)}.$$

Two cases may arise.

– *Case* $|A| \geq n_l^{\beta}$. Since there are at least $|A|$ edges of length $\frac{r}{h}$, it holds that

$$\mathsf{cost}(T) \geq \frac{r}{h} \cdot |A| = \Omega(l \cdot n_l^{\beta+\frac{1}{d}}) = \Omega\left(l \cdot n_l^{1+\frac{1}{g(h)}}\right).$$

– *Case* $|A| < n_l^{\beta}$. For every point x in B there is a path from x to the root s with at most h hops. Since the distance from x to s is at least r, then in the path there is at least one edge of length at least r/h. Hence, we can partition the points in $A \cup B$ into $|A|$ subsets $A_1, A_2, \ldots A_{|A|}$ where a point y is in A_i if a_i is the first point in A in the path from y to s. Notice that the points in the subsets A_i, $1 \leq i \leq |A|$, belong to (edge-)disjoint subtrees $T_1, T_2, \ldots, T_{|A|}$ of T where T_i is an $(h-1)$-tree rooted at a_i. Let $n_{l,i}$ be the number of d-cubes containing the points of T_i, $1 \leq i \leq |A|$. It holds that

$$\begin{aligned} \mathsf{cost}(T) &\geq \sum_{i=1}^{|A|} \mathsf{cost}(T_i) \\ &= \Omega\left(\sum_{i=1}^{|A|} l \cdot n_{l,i}^{\;1+\frac{1}{g(h-1)}}\right) \quad \text{by inductive hypothesis} \end{aligned}$$

$$
\begin{aligned}
&= \Omega\left(l \cdot |A| \cdot \left(\frac{\sum_{i=1}^{|A|} n_{l,i}}{|A|}\right)^{1+\frac{1}{g(h-1)}}\right) && \text{by the Hölder inequality} \\
&= \Omega\left(l \cdot |A|^{-\frac{1}{g(h-1)}} \cdot {n_l}^{1+\frac{1}{g(h-1)}}\right) && \text{since } \sum_{i=1}^{|A|} n_{l,i} \geq n'_l \geq \frac{n_l}{2} \\
&= \Omega\left(l \cdot n_l^{-\frac{\beta}{g(h-1)}+1+\frac{1}{g(h-1)}}\right) && \text{since } |A| < n_l^{\beta} \\
&= \Omega\left(l \cdot n_l^{1+\frac{g(h)-d}{d \cdot g(h-1) \cdot g(h)}}\right) \\
&= \Omega\left(l \cdot n_l^{1+\frac{d \cdot g(h-1)}{d \cdot g(h-1) \cdot g(h)}}\right) && \text{since } g(h) = d \cdot g(h-1) + d \\
&= \Omega\left(l \cdot n_l^{1+\frac{1}{g(h)}}\right).
\end{aligned}
$$

The thesis follows. □

By applying the probabilistic method of *bounded differences* [19], we can prove Theorem 1.

Proof of Theorem 1. Let us partition the d-cube into n d-cubes, each of them of side length $l = Ln^{-\frac{1}{d}}$. Let n_l be the number of nonempty d-cubes. Lemma 3 implies that

$$
\mathsf{cost}(T) = \begin{cases} \Omega\left(L \cdot n^{-1} \cdot n_l^{1+\frac{1}{h}}\right) & \text{if } d = 1 \\ \Omega\left(L \cdot n^{-\frac{1}{d}} \cdot n_l^{1+\frac{d-1}{d^{h+1}-d}}\right) & \text{otherwise.} \end{cases}
$$

The theorem follows by noticing that, by applying the method of *bounded differences* [19], we have $n_l \geq n/4$, with high probability. □

4. The divide et Impera heuristic

The h-PARTY heuristic is described in Algorithm 1. Observe that the value of k determines the number of stations (with their costs) that are directly connected to the root station. As we have already mentioned in the introduction, the choice of a "good" k is the main technical problem: our solution forces the cost due to the stations directly connected to the root in order to match the lower bound given in Lemma 3. The details of this argument are explained in the proof of the Theorem 2.

Proof of Theorem 2. We equivalently show that

$$
\mathsf{cost}(T) = O\left(L \cdot n^{1-\frac{1}{d}+\frac{1}{g(h)}}\right)
$$

where $g(h)$ is the function introduced in Lemma 3, that is

$$
g(h) = \begin{cases} d & \text{if } h = 1 \\ d \cdot g(h-1) + d & \text{otherwise.} \end{cases}
$$

The proof is by induction on h. If $h = 1$ it is clear that $\mathsf{cost}(T) = O(L \cdot n)$.

For $h \geq 2$, let t be the number of nonempty d-cubes in the d-cube of size length L and $\{q_1, q_2, \ldots q_t\}$ be the set of points selected by the procedure in the t non-empty d-cubes; let T_i be the $(h-1)$-tree rooted in q_i and S_i be the set of points spanned by T_i, $1 \leq i \leq t$. By inductive hypothesis, we get $\mathsf{cost}(T_i) = O\left(\frac{L}{k^{\frac{1}{d}}} \cdot |S_i|^{1-\frac{1}{d}+\frac{1}{g(h-1)}}\right)$. We thus

Algorithm 1 h-PARTY(S, s)

```
if h = 1 then
  T ← {{x, s}|x ∈ S − {s}};
else
  T ← ∅;
  if d = 1 then
    k ← ⌊|S|^(1/h)⌋;
  else
    k ← ⌊|S|^(1 − 1/d + (d−1)/(d^(h+1) − d))⌋; {this choice is explained in the text}
  end if
  Let L be the side length of the smallest d-cube containing all points in S;
  Partition the d-cube into d-cubes of side length L/⌈k^(1/d)⌉;
  Let k' be the number of d-cubes and let S_i be the points of S in the i-th d-cube, 1 ≤ i ≤ k';
  for i ← 1 to k' do
    if |S_i| ≥ 1 then
      choose a point s' in S_i;
      T ← T ∪ {{s', s}};
      if |S_i| > 1 then
        T ← T∪(h − 1)-PARTY(S_i, s');
      end if
    end if
  end for
end if
return T
```

have that

$$
\begin{aligned}
\mathrm{cost}(T) &= \sum_{i=1}^{t} d(q_i, s) + \sum_{i=1}^{t} \mathrm{cost}(T_i) \\
&\leq L \cdot t + \sum_{i=1}^{t} \mathrm{cost}(T_i) && \text{since } d(q_i, s) \leq L \\
&= O\left(L \cdot t + \sum_{i=1}^{t} \frac{L}{k^{\frac{1}{d}}} \cdot |S_i|^{1-\frac{1}{d}+\frac{1}{g(h-1)}}\right) && \text{by inductive hyp.} \\
&= O\left(L \cdot t + \frac{L}{k^{\frac{1}{d}}} \cdot t \cdot \left(\frac{\sum_{i=1}^{t} |S_i|}{t}\right)^{1-\frac{1}{d}+\frac{1}{g(h-1)}}\right) && \text{by Hölder ineq.} \\
&= O\left(L \cdot t + \frac{L}{k^{\frac{1}{d}}} \cdot t^{\frac{1}{d}-\frac{1}{g(h-1)}} \cdot n^{1-\frac{1}{d}+\frac{1}{g(h-1)}}\right) && \text{since } \sum_{i=1}^{t} |S_i| = n \\
&= O\left(L \cdot k + L \cdot k^{-\frac{1}{g(h-1)}} \cdot n^{1-\frac{1}{d}+\frac{1}{g(h-1)}}\right) && \text{since } t \leq k \\
&= O\left(L \cdot n^{1-\frac{1}{d}+\frac{1}{g(h)}} + L \cdot n^{-\frac{1}{g(h-1)}(1-\frac{1}{d}+\frac{1}{g(h)})+1-\frac{1}{d}+\frac{1}{g(h-1)}}\right) \\
&= O\left(L \cdot n^{1-\frac{1}{d}+\frac{1}{g(h)}} + L \cdot n^{1-\frac{1}{d}+\frac{g(h)-d}{d \cdot g(h-1) \cdot g(h)}}\right) \\
&= O\left(L \cdot n^{1-\frac{1}{d}+\frac{1}{g(h)}}\right).
\end{aligned}
$$

where the last step follows since

$$\frac{g(h)-d}{d\cdot g(h-1)\cdot g(h)}=\frac{d\cdot g(h-1)}{d\cdot g(h-1)\cdot g(h)}=\frac{1}{g(h)}.\quad\square$$

Finally, it is not hard to verify that, for any $h > 0$, the worst-case time complexity is $O(n)$.

5. Conclusions and open problems

In this paper, we have provided the first nontrivial lower bound on the solution cost *for almost all* Euclidean instances (i.e. the lower bound holds with high probability). Then, we have introduced an easy-to-implement, fast divide and impera heuristic whose solution-cost matches the lower bound. We finally remark that the proof of Lemma 3 strongly relies on the fact that h and d do not depend on n. It thus follows that an interesting future work consists in extending our asymptotical analysis to nonconstant h (e.g. $h = \Omega(\log n)$).

References

[1] L. Alfandari, V.T. Paschos, Approximating minimum spanning tree of depth 2, International Transactions in Operations Research 6 (1999) 607–622.
[2] E. Althaus, S. Funke, S. Har-Peled, J. Koenemann, E.A. Ramos, M. Skutella, Approximation k-hop minimum-spanning trees, Operations Research Letters 33 (2005) 115–120.
[3] S. Arora, P. Raghavan, S. Rao, Approximation schemes for Euclidean k-medians and related problems, in: Proc. 30-th ACM Symposium on Theory of Computing, 1998, pp. 106–113.
[4] K. Bala, K. Petropoulos, T.E. Stern, Multicasting in a linear lightwave network, in: Proc. of INFOCOM, 1993, pp. 1350–1358.
[5] A. Balakrishnan, K. Altinkemer, Using a hop-constrained model to generate alternative communication network design, ORSA Journal of Computing 4 (1992) 147–159.
[6] A. Bookstein, S.T. Klein, Compression of correlated bit-vectors, Information Systems 16 (4) (1996) 110–118.
[7] R. Chou, T. Johnson, Distributed Operating Systems and Algorithms, Addison-Wesley, Reading, MA, 1997.
[8] F.A. Chudak, Improved approximation algorithms for uncapacitated facility location problem, in: Proc. of the 6-th Conference on Integer Programming and Combinatorial Optimization, 1998.
[9] A.E.F. Clementi, M. Di Ianni, A. Monti, G. Rossi, R. Silvestri, Experimental analysis of practically efficient algorithms for bounded-hop accumulation in ad-hoc wireless networks, in: Proc. of the IEEE IPDPS-WMAN, 2005.
[10] A. Ephremides, G.D. Nguyen, J.E. Wieselthier, On the construction of energy-efficient broadcast and multicast trees in wireless networks, in: Proc. of the 19th INFOCOM, 2000, pp. 585–594.
[11] L. Gouveia, Using the Miller–Tucker–Zemlin constraints to formulate a minimal spanning tree problem with hop constraints, Computers and Operations Research 22 (1995) 959–970.
[12] L. Gouveia, Multicommodity flow models for spanning trees with hop constraints, European Journal of Operational Research 95 (1996) 178–190.
[13] L. Gouveia, C. Requejo, A new relaxation approach for the hop-constrain minimum spanning tree problem, European Journal of Operational Research 132 (2001) 539–552.
[14] S. Guha, S. Khuller, Greedy strikes back: Improved facility location algorithms, Journal of Algorithms 31 (1999) 228–248.
[15] M. Haenggi, Twelve reasons not to route over many short hops, in: Proc. IEEE Vehicular Technology Conference, VTC'04 Fall, Los Angeles, CA, September 2004.
[16] G. Kortsarz, D. Peleg, Approximating the weight of shallow Steiner trees, Discrete Applied Mathematics 93 (1999) 265–285.
[17] M.R. Korupolu, C.G. Plaxton, R. Rajaraman, Analysis of a local search heuristic for facility location problems, in: Proc. of the 9th annual ACM-SIAM Symposium on Discrete algorithms, 1998, pp. 1–10.
[18] M. Mahdian, Y. Ye, J. Zhang, A 1.52-approximation algorithm for the uncapacitated facility location problem, in: Proc. of APPROX 2002, in: LNCS, vol. 2462, 2002, pp. 229–242.
[19] C.J.H. McDiarmid, On the method of bounded differences, in: J. Siemons (Ed.), Surveys in Combinatorics: Invited Papers at the 12th British Combinatorial Conference, in: Lecture Notes Series, vol. 141, London Mathematical Society, 1989, pp. 148–188.
[20] G.R. Raidl, B.A. Julstrom, Greedy heuristics and an evolutionary algorithm for the bounded-diameter minimum spanning tree problem, in: Proc. of the 2003 ACM Symposium on Applied Computing 2003, pp. 747–752.
[21] K. Raymond, A tree-based algorithm for distributed mutual exclusion, ACM Transactions on Computer Systems 7 (1) (1989) 61–77.
[22] D.B. Shmoys, E. Tardos, K. Aardal, Approximation algorithms for facility location problems, in: Proc. of the 29-th Annual ACM Symposium on Theory of Computing, 1997, pp. 265–274.
[23] S. Voss, The steiner tree problem with hop constraint, Annals of Operations Research 86 (1999) 321–345.