

INSTITUT FÜR INFORMATIK

**Randomized Algorithms for Mixed
Matching and Covering in Hypergraphs
in 3D Seed Reconstruction in
Brachytherapy**

Helena Fohlin, Lasse Kliemann, and Anand
Srivastav

Bericht Nr. 0611

Juli 2006



CHRISTIAN-ALBRECHTS-UNIVERSITÄT

KIEL

Institut für Informatik der
Christian-Albrechts-Universität zu Kiel
Olshausenstr. 40
D – 24098 Kiel

**Randomized Algorithms for Mixed Matching and
Covering in Hypergraphs in 3D Seed
Reconstruction in Brachytherapy**

Helena Fohlin, Lasse Kliemann, and Anand Srivastav

Bericht Nr. 0611
Juli 2006

e-mail: {hef,lki,asr}@numerik.uni-kiel.de

Randomized Algorithms for Mixed Matching and Covering in Hypergraphs in 3D Seed Reconstruction in Brachytherapy

Helena Fohlin, Lasse Kliemann*, and Anand Srivastav

1 Introduction

Brachytherapy is a method developed in the 1980s for cancer radiation in organs like prostate, lung, or breast. At the Clinic of Radiotherapy (radiooncology), Christian Albrechts University of Kiel, among others, low dose radiation therapy (LDR therapy) for the treatment of prostate cancer is applied, where 25-80 small radioactive seeds are implanted in the affected organ. They have to be placed so that the tumor is exposed with sufficiently high radiation and adjacent healthy tissue is exposed to a radiation dose as low as possible. Unavoidably, the seeds can move due to blood circulation, movements of the organ, etc. For the quality control of the treatment plan the locations of the seeds after the operation have to be checked. This is done by taking usually 3 X-ray photographs from three different angles (so-called 3-film technique). On the films the seeds appear as white lines. To determine the positions of the seeds in the organ the task now is to match the 3 different images (lines) representing the same seed.

Previous Work. The 3-film technique was independently applied by Rosenthal and Nath [21], Biggs and Kelley [10] and Altschuler, Findlay, Epperson [3], while Siddon and Chin [13] applied a special 2-film technique that took the seed endpoints as image points rather than the seed centers. The algorithms invoked in these papers are matching heuristics justified by experimental results. New algorithmic efforts were taken in the last 5 years. Tubic, Zaccarin, Beaulieu and Pouliot [9] used simulated annealing, Todor, Cohen, Amols and Zaider [5] combined several heuristic approaches, and Lam, Cho, Marks and Narayanan [15] introduced the so-called Hough transform, a standard method in image processing and computer vision for the seed reconstruction problem. Recently, Narayanan, Cho and Marks [14] addressed also the problem of reconstruction with incomplete data set. These papers essentially focus on improvement of the geometric projections. From the mathematical programming side branch-and-bound was applied by Balas and Saltzman [8] and Brogan [11]. These papers provide the link to integer programming models of the problem.

*Supported by DFG.

None of these papers give a mathematical analysis or provable performance guarantee of the algorithms in use. In particular, since different projection techniques essentially result in different objective functions, it would be desirable to have an algorithm which is independent of the specific projection technique and thus is applicable to all such situations. Furthermore, it is today considered as a challenging task in algorithmic discrete mathematics and theoretical computer science to give fast algorithms for NP -hard problems, which provably (or at least in practice) approximate the optimal solution. This is sometimes a fast alternative to branch-and-bound methods.

Our Contribution. In this paper we model the seed reconstruction problem as a minimum-weight perfect matching problem in a hypergraph: we consider a complete 3-uniform hypergraph, where its nodes are the seed images on the 3 films, and each of its hyperedges contains 3 nodes (one from each X-ray photo). We define a weight function for the hyperedges, which is close to zero if the three lines from a hyperedge belong to the same seed and increases otherwise. The goal is to find a matching, i.e., a subset of pairwise disjoint hyperedges, so that all nodes are covered and the total weight of these hyperedges is minimum. This is nothing else than the minimum-weight perfect matching problem in a hypergraph. Since this problem generalizes the NP -hard 3-dimensional assignment problem (see [16]), it is NP -hard as well. Thus we can only hope to find an algorithm which solves the problem approximately in polynomial time, unless $P = NP$.

We model the problem as an integer linear program. To solve this integer program, an algorithm based on the so-called randomized rounding scheme introduced by Raghavan and Thompson [23] is designed and applied. This algorithm is not only very fast, but accessible at least in part for a mathematical rigorous analysis. We give a partial analysis of the algorithm combining probabilistic and combinatorial methods, which shows that in the worst-case the solution produced is in some strong sense close to a minimum-weight perfect matching. The heart of the analytical methods are tools from probability theory, like large deviation inequalities. All in all, our algorithm points towards a mathematical rigorous analysis of heuristics for the seed reconstruction problem and is practical as well. Furthermore, the techniques developed here are promising for an analysis of mixed integer packing and covering problems, which are of independent interest in discrete optimization. Moreover, we show that an implementation of our algorithm is very effective on a set of patient data from the Clinic of Radiotherapy, Christian Albrechts University of Kiel. In fact, the algorithm for a certain choice of parameters outputs optimal or nearly optimal solutions where only a few seeds are unmatched. It is interesting that the practical results are much better than the results of the theoretical analysis do indicate. Here we have the challenging situation of closing the gap between the theoretical analysis and the good practical performance, which should be addressed in future work.

In conclusion, while in previous work on the seed reconstruction problem only heuristics were used, this paper is a first step in designing mathematical analyzable *and* practically efficient algorithms.

The paper is organized as follows.

In Section 2 we describe the seed reconstruction problem more precisely and give a mathematical model. For this we introduce the notion of (b, k) -matching which generalizes the notions of b -matching in hypergraphs and partial k -covering in hypergraphs. In fact, a (b, k) -matching is a b -matching, i.e., a subset of hyperedges such that no node is incident in more than b of them, covering at least k nodes. So for a hypergraph with n nodes, $(1, n)$ -matching is a perfect matching problem. Furthermore, some large deviation inequalities are listed as well.

In Section 3 we give an integer linear programming formulation for the (b, k) -matching problem and state the randomized rounding algorithm. This algorithm solves the linear programming (LP) relaxation up to optimality and then generates an integer solution by picking edges with the probabilities given by the optimal LP-solution. After this procedure we remove edges in a greedy way to get a feasible b -matching.

In Section 4 we analyze the algorithm with probabilistic tools.

In Section 5 we test the practical performance of the algorithm on real patient data for five patients treated in the Clinic of Radiotherapy in Kiel. The algorithm is implemented in C++, and is iterated for each patient data set 100 times. For most of the patients all seeds are matched if we choose good values of the parameters, i.e., letting them be close to the values enforcing a minimum-weight perfect matching. The algorithm is very fast: within a few seconds of CPU time on a PC, it delivers the solution.

2 Hypergraph Matching Model of 3D Seed Reconstruction

Brachytherapy is a cancer radiation therapy developed in the 1980s. In the low dose variant of brachytherapy, about 25 to 80 small radioactive implants called seeds are placed in organs like prostate, lung or breast, and remain there. A seed is a titan cylinder of length approx. 4.5 mm encapsulating radioactive material like Iod-125 or Pd-103.

The method allows an effective continuous radiation of tumor tissue with a relatively low dose for a long time in which radiation is delivered at a very short distance to the tumor by placing the radioactive source in the affected organ. Today it is a widely spread technique and an alternative to the usual external radiation. A benefit for the patient is certainly that he/she has not to suffer from a long treatment with various radiation sessions. For the treatment of prostate cancer brachytherapy has been reported as a very effective method [7]. At the Clinic of Radiotherapy, Christian Albrechts University of Kiel, brachytherapy has become the standard radiation treatment of prostate cancer.

2.1 The Optimization Problem

In LDR brachytherapy with seeds two mathematical optimization problems play a central role:

a) Placement Problem

The most important problem is to determine a minimum number of seeds along with their placement in the organ. The placement must be such that

- a) the tumor tissue is exposed with sufficient dose avoiding cold spots (regions with insufficient radiation) and hot spots (regions with too much radiation) and
- b) normal tissue or critical regions like the urethra are exposed with a minimum possible, medical tolerable dose.

The problem thus appears as a combination of several NP -hard multi-criteria optimization problems, e.g., set covering and facility location with restricted areas. Since the dose distribution emitted by the seeds is highly nonlinear, the problem is further complicated beyond set covering with regular geometric objects, like balls, ellipsoids, etc. An intense research has been done in this area in the last 10 years. Among the most effective placement tools are the mixed-integer programming methods proposed by E. Lee [20]. At the Clinic of Radiotherapy, Christian Albrechts University of Kiel, a commercial placement software (VariSeed[®] of the company VARIAN) is applied. The software offers a two-film and a three-film technique. According to the manual of the software, the three-film technique is an ad hoc extension of the two-film technique of Chin and Siddon [13].

b) 3D Seed Reconstruction Problem

After the operative implantation of the seeds, due to blood circulation and movements of the organ, resp. patient, the seeds can change their original position. Usually 1-2 hours after the operation a determination of the actual seed positions in the organ is necessary in order to control the quality and to take further steps. In the worst case a short high dose radiation (HDR brachytherapy) has to be conducted.

The seed locations are determined by 3 X-ray films of the organ taken from 3 different angles, see Figures 2, 3, and 4. This technique was introduced by Amols and Rosen [4] in 1981. The advantage of the 3-film technique compared with the 2-film technique is that it seems to be less ambiguous in identifying seed locations. So, each film shows the seeds from a different 3-dimensional perspective. The task is to determine the location of the seeds in the organ by matching seed images on the 3 films.

To formalize the seed reconstruction problem, an appropriate geometrical measure as a cost function for matching 3 seed images from each film is introduced. We show in the following how the cost function is computed for the upper endpoint of the seed, see Figure 1. The cost of the lower endpoint is calculated in the same way.

For the three seed images we have three lines P_1, P_2, P_3 connecting the lower resp. upper endpoint of the seed images with the X-ray source. We determine the shortest connections between the lines P_i and P_j for all i, j . Let $r_i = (x_i, y_i, z_i)$ be the centers of the shortest connections and let $\bar{x}, \bar{y}, \bar{z}$ be the mean values of the x, y, z coordinate of r_1, r_2, r_3 . We define the standard deviation

$$\Delta r = \sqrt{\frac{1}{3} \sum_{i=1}^3 (x_i - \bar{x})^2} + \sqrt{\frac{1}{3} \sum_{i=1}^3 (y_i - \bar{y})^2} + \sqrt{\frac{1}{3} \sum_{i=1}^3 (z_i - \bar{z})^2}.$$

The cost for the upper (resp. lower) endpoint of *any* choice of 3 seed images from the three X-ray photos is the Δr of the associated lines. It is clear that Δr is close to zero if the 3 seed images represent the same seed.

The *total cost* for three seed images is the sum of the standard deviation Δr for the upper endpoint and the standard deviation for the lower endpoint.¹

An alternative cost measure can be the area spanned by the triangle r_1, r_2, r_3 . But in this paper this cost function is not considered.

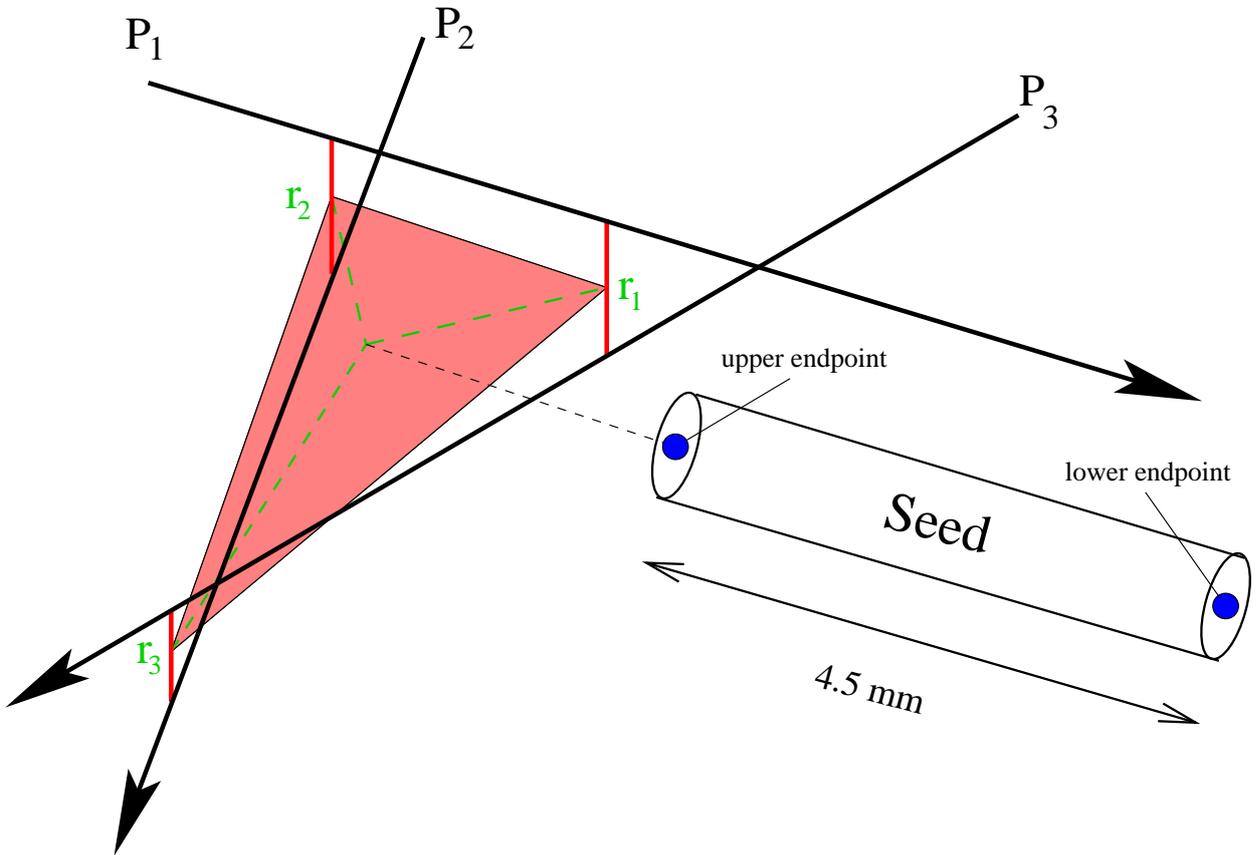


Figure 1: cost function for the upper endpoint

If the cost function is well posed, the optimal solution of the problem should be in one-to-one correspondence to the real seed locations in the organ. Thus the problem reduces to

¹By appropriate scaling Δr to $\Delta r/\alpha$, with some $\alpha \geq 1$, one can assume that the total cost is in $[0, 1]$.

a three-dimensional assignment (or matching) problem, where we minimize the cost of the matching. In the literature the problem is also noted as the AP3 problem, which is NP -hard. Thus under the hypothesis $P \neq NP$, we cannot expect an efficient, i.e., polynomial time algorithm solving the problem to optimality.

2.2 Hypergraph Matching and Seed Reconstruction

We use the standard notion of graphs and hypergraphs. A finite graph $G = (V, E)$ is a pair of a finite set V (the set of vertices or nodes) and a subset $E \subseteq \binom{V}{2}$, where $\binom{V}{2}$ denotes the set of all 2-element subsets of V . The elements of E are called edges. A hypergraph (or set system) $\mathcal{H} = (V, \mathcal{E})$ is a pair of a finite set V and a subset \mathcal{E} of the power set $\mathcal{P}(V)$. The elements of \mathcal{E} are called hyperedges.

Let $\mathcal{H} = (V, \mathcal{E})$ be a hypergraph. For $v \in V$ we define

$$\deg(v) := |\{E \in \mathcal{E}; v \in E\}| \quad \text{and} \quad \Delta = \Delta(\mathcal{H}) := \max_{v \in V} \deg(v).$$

We call $\deg(v)$ the vertex-degree of v and $\Delta(\mathcal{H})$ is the maximum vertex degree of \mathcal{H} . The hypergraph \mathcal{H} is called r -regular resp. s -uniform, if $\deg(v) = r$ for all $v \in V$ resp. $|E| = s$ for all $E \in \mathcal{E}$. It is convenient to order the vertices and hyperedges, $V = \{v_1, \dots, v_n\}$ and $\mathcal{E} = \{E_1, \dots, E_m\}$, and to identify vertices and edges with their indices. The hyperedge-vertex incidence matrix of a hypergraph $\mathcal{H} = (V, \mathcal{E})$, with $V = \{v_1, \dots, v_n\}$ and $\mathcal{E} = \{E_1, \dots, E_m\}$, is the matrix $A = (a_{ij}) \in \{0, 1\}^{m \times n}$, where $a_{ij} = 1$ if $v_j \in E_i$, and 0 else. Sometimes the vertex-hyperedge incidence matrix A^T is used.

We proceed to the formulation of a mathematical model for the seed reconstruction problem.

Definition 2.1. *Let $\mathcal{H} = (V, \mathcal{E})$ be a hypergraph with $|V| = n, |\mathcal{E}| = m$. Let $w : \mathcal{E} \rightarrow \mathbb{Q} \cap [0, 1]$ be a weight function. Let $b, k \in \mathbb{N}$.*

- (i) *A b -matching in \mathcal{H} is a subset $\mathcal{E}^* \subseteq \mathcal{E}$ such that each $v \in V$ is contained in at most b edges of \mathcal{E}^* .*
- (ii) *A (b, k) -matching \mathcal{E}^* is a b -matching, such that at least k vertices are covered by edges of \mathcal{E}^* .*
- (iii) *For a subset $\mathcal{E}^* \subseteq \mathcal{E}$, we define its weight $w(\mathcal{E}^*)$ as the sum of the weights of the edges from \mathcal{E}^* .*

We consider the following optimization problem.

Problem 2.2. Min- (b, k) -Matching:

Find a (b, k) -matching with minimum weight, if such a matching exists.

This problem, for certain choices of b and k , specializes to well-known problems in combinatorial optimization:

1. Min-(1, n)-Matching is the minimum-weight perfect matching problem in hypergraphs.
2. Min-(m, n)-Matching is the set covering problem in hypergraphs
3. Min-(m, k)-Matching is the partial set covering (or k -set covering) problem in hypergraphs.

The seed reconstruction problem can be modeled as a minimum-weight perfect matching problem in a 3-uniform hypergraph as follows: let V_1, V_2, V_3 be the seed images on the X-ray photos 1, 2, 3. With $V = V_1 \cup V_2 \cup V_3$ and $\mathcal{E} = V_1 \times V_2 \times V_3$ the hypergraph under consideration is $\mathcal{H}=(V, \mathcal{E})$. Given a weight function $w : \mathcal{E} \rightarrow \mathbb{Q} \cap [0, 1]$, the seed reconstruction problem is just the problem of finding the minimum weight perfect matching in \mathcal{H} .

2.3 Some Probabilistic Tools

Throughout this article we consider only finite probability spaces (Ω, \mathbb{P}) , where Ω is a finite set and \mathbb{P} is a probability measure with respect to the power set $\mathcal{P}(\Omega)$ as the sigma field. We recall the basic Markov and Chebyshev inequalities.

Theorem 2.3. (Markov Inequality) *Let (Ω, \mathbb{P}) be a probability space and $X : \Omega \rightarrow \mathbb{R}^+$ a random variable with expectation $\mathbb{E}(X) < \infty$. Then for any $\lambda \in \mathbb{R}^+$*

$$\mathbb{P}[X \geq \lambda] \leq \frac{\mathbb{E}(X)}{\lambda}.$$

An often sharper bound is the well-known inequality of Chebyshev:

Theorem 2.4. (Chebyshev Inequality) *Let (Ω, \mathbb{P}) be a probability space and $X : \Omega \rightarrow \mathbb{R}$ a random variable with finite expectation $\mathbb{E}(X)$ and variance $\text{Var}(X)$. Then for any $\lambda \in \mathbb{R}^+$*

$$\mathbb{P}[|X - \mathbb{E}(X)| \geq \lambda\sqrt{\text{Var}(X)}] \leq \frac{1}{\lambda^2}.$$

For one-sided deviation the following Chebyshev-Cantelli inequality (see [2]) gives better bounds:

Theorem 2.5. *Let X be a non-negative random variable with finite expectation $\mathbb{E}(X)$ and variance $\text{Var}(X)$. Then for any $a > 0$*

$$\mathbb{P}[X \leq \mathbb{E}(X) - a] \leq \frac{\text{Var}(X)}{\text{Var}(X) + a^2}.$$

The following estimate on the variance of a sum of dependent random variables can be proved as in [2] Corollary 4.3.3.

Let X be the sum of any 0/1 random variables, i.e., $X = X_1 + \dots + X_n$, and let $p_i = \mathbb{E}(X_i)$ for all $i = 1, \dots, n$. For a pair $i, j \in \{1, \dots, n\}$ we write $i \sim j$, if X_i and X_j are dependent. Let Γ be the set of all unordered dependent pairs i, j , i.e., the 2-element sets $\{i, j\}$, and let

$$\gamma = \sum_{\{i,j\} \in \Gamma} \mathbb{E}(X_i X_j)$$

Proposition 2.6.

$$\text{Var}(X) \leq \mathbb{E}(X) + 2\gamma$$

Proof: We have

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}[X_i, X_j], \quad (1)$$

where the second sum is over ordered pairs.

Since $X_i^2 = X_i$, and $\text{Var}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \mathbb{E}(X_i)(1 - \mathbb{E}(X_i)) \leq \mathbb{E}(X_i)$, (1) gives

$$\text{Var}(X) \leq \mathbb{E}(X) + \sum_{i \neq j} \text{Cov}[X_i, X_j] \quad (2)$$

If $i \not\sim j$, then $\text{Cov}[X_i, X_j] = 0$. For $i \sim j$ we have

$$\text{Cov}[X_i, X_j] = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) \leq \mathbb{E}(X_i X_j), \quad (3)$$

so (3) implies the assertion of the proposition. \square

We proceed to the statement of standard large deviation inequalities for a sum of independent random variables.

Let X_1, \dots, X_n be 0/1 valued mutually independent (briefly: independent) random variables, where

$$\mathbb{P}[X_j = 1] = p_j, \quad \mathbb{P}[X_j = 0] = 1 - p_j$$

for probabilities $p_j \in [0, 1]$ for all $1 \leq j \leq n$. For $1 \leq j \leq n$ let w_j denote rational weights with

$$0 \leq w_j \leq 1$$

and let

$$X = \sum_{j=1}^n w_j X_j.$$

The sum

$$X = \sum_{j=1}^n w_j X_j \quad \text{with } w_j = 1 \quad \forall j \in \{1, \dots, n\} \quad (4)$$

is the well-known binomially distributed random variable with mean np . The inequalities given below can be found in the books of Alon, Spencer and Erdős [2], Habib, McDiarmid, Ramirez-Alfonsin and Reed [17], and Janson, Łuczak, Ruciński [19].

The following basic large deviation inequality is implicitly given in Chernoff [12] in the binomial case. In explicit form it can be found in Okamoto [22]. Its generalization to arbitrary weight is due to Hoeffding [18].

Theorem 2.7. (Hoeffding 1963) *Let $\lambda > 0$ and let X be as in (4). Then*

$$(a) \mathbb{P}(X > \mathbb{E}(X) + \lambda) \leq e^{-\frac{2\lambda^2}{n}}$$

$$(b) \mathbb{P}(X < \mathbb{E}(X) - \lambda) \leq e^{-\frac{2\lambda^2}{n}}.$$

In the literature Theorem 2.7 is well known as the Chernoff bound. For small expectations, i.e $\mathbb{E}(X) \leq \frac{n}{6}$, the following inequalities due to Angluin and Valiant [6] give better bounds.

Theorem 2.8. *Let X_1, \dots, X_n be independent random variables with $0 \leq X_i \leq 1$ and $\mathbb{E}(X_i) = p_i$ for all $i = 1, \dots, n$. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}(X)$. For any $\beta > 0$*

$$(i) \mathbb{P}[X \geq (1 + \beta) \cdot \mu] \leq e^{-\frac{\beta^2 \mu}{2(1+\beta/3)}}$$

$$(ii) \mathbb{P}[X \leq (1 - \beta) \cdot \mu] \leq e^{-\frac{\beta^2 \mu}{2}}$$

Note that for $0 \leq \beta \leq 3/2$ the bound in (i) is at most $\exp(-\beta^2 \mu/3)$.

We will also need the Landau symbols O, o, Θ and Ω .

Definition 2.9. *Let $f : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}, g : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ be functions. Then*

- $f(n) = O(g(n))$ if $\exists c_1, c_2 \in \mathbb{R}_{>0}$, such that

$$f(n) \leq c_1 g(n) + c_2 \quad \text{for all } n \in \mathbb{N}.$$

- $f(n) = \Omega(g(n))$ if $g(n) = O(f(n))$.
- $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.
- $f(n) = o(g(n))$ if $\frac{f(n)}{g(n)} \xrightarrow{n \rightarrow \infty} 0$ (provided that $g(n) \neq 0$ for all n large enough).

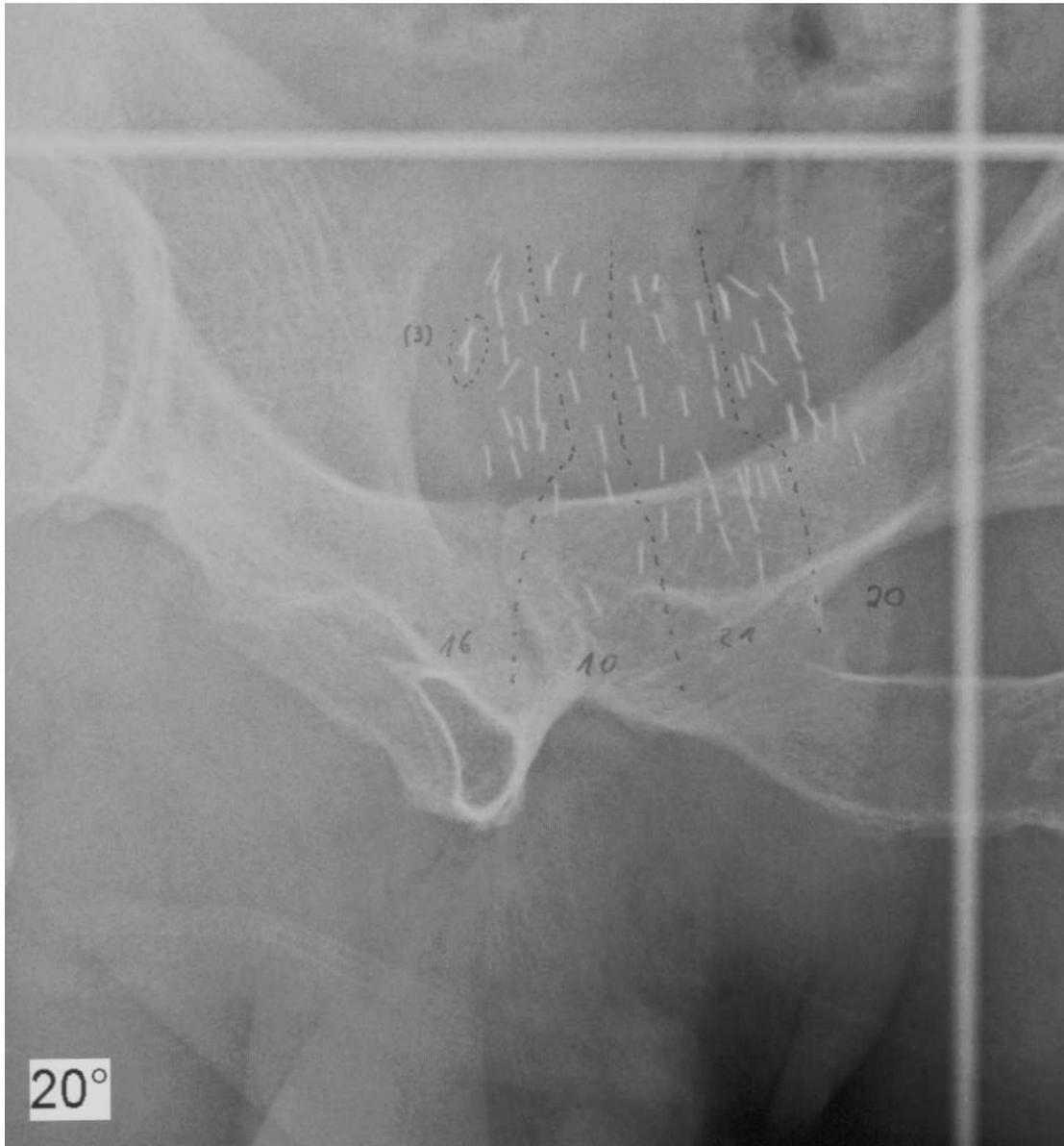


Figure 3: X-ray, 20 degrees

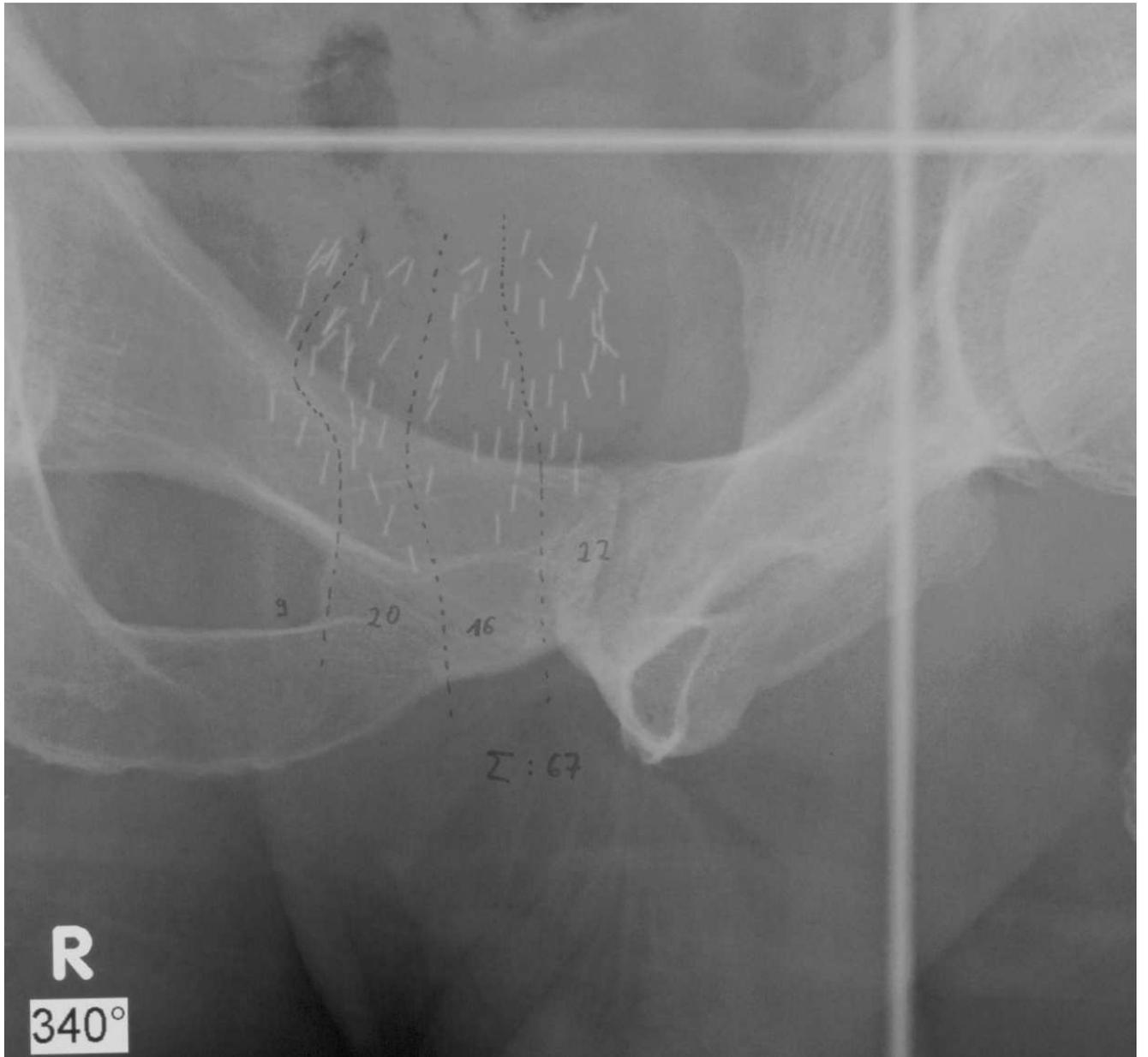


Figure 4: X-ray, 340 degrees

3 Simultaneous Matching and Covering Algorithms

In this section we present a randomized algorithm for the (b, k) -matching problem.

3.1 Randomized Algorithms for (b, k) -Matching

Let $\mathcal{H} = (V, \mathcal{E})$, $|V| = n$, $|\mathcal{E}| = m$ be a hypergraph. We identify the nodes resp. edges of \mathcal{H} by their indices, so $V = \{1, \dots, n\}$ and $\mathcal{E} = \{1, \dots, m\}$. Let $b \geq 1$.

An integer programming formulation of the minimum-weight (b, k) -matching is the following:

$$\begin{aligned} \text{Min-}(b, k)\text{-ILP} \quad & \min \sum_{i=1}^m w_i X_i \\ & \sum_{i=1}^m a_{ij} X_i \leq b \quad \forall j \in \{1, \dots, n\} \end{aligned} \tag{5}$$

$$\sum_{i=1}^m a_{ij} X_i \geq Y_j \quad \forall j \in \{1, \dots, n\} \tag{6}$$

$$\sum_{j=1}^n Y_j \geq k \tag{7}$$

$$X_i, Y_j \in \{0, 1\} \quad \forall i \in \{1, \dots, m\} \text{ and } \forall j \in \{1, \dots, n\} \tag{8}$$

Note that Min- (b, n) -ILP is equivalent to the minimum-weight perfect b -matching problem and Min- (b, k) -ILP is a b -matching problem with a k -partial covering of the vertices.

For the minimum-weight perfect b -matching problems in hypergraphs, where a perfect b -matching exists, for example the 3-uniform hypergraph associated to the seed reconstruction problem an alternative integer linear programming formulation using local covering conditions is useful.

We add the condition $\sum_{i=1}^m a_{ij} X_i \geq \theta$ for some $\theta \in (0, 1]$ for all $j \in \{1, \dots, n\}$ to Min- (b, k) -ILP. Then, by integrality all vertices are covered and any feasible solution of such an ILP is a perfect b -matching. For the integer program the additional condition is redundant, but since the LP-relaxation of Min- (b, k) -ILP together with the inequality has a smaller feasible region than the LP-relaxation of Min- (b, k) -ILP, the gap between the integer optimum and the feasible LP-optimum might be smaller as well. This leads to a better ‘‘approximation’’ of the integer optimum by the LP-optimum.

Furthermore we will see in the theoretical analysis (Section 4) that we can cover significantly more nodes if we add this condition.

$$\begin{aligned}
\text{Min-}(b, k, \theta)\text{-ILP} \quad & \min \sum_{i=1}^m w_i X_i \\
& \sum_{i=1}^m a_{ij} X_i \leq b \quad \forall j \in \{1, \dots, n\} & (9) \\
& \sum_{i=1}^m a_{ij} X_i \geq Y_j \quad \forall j \in \{1, \dots, n\} & (10) \\
& \sum_{i=1}^m a_{ij} X_i \geq \theta \quad \forall j \in \{1, \dots, n\} & (11) \\
& \sum_{j=1}^n Y_j \geq k & (12) \\
& X_i, Y_j \in \{0, 1\} \quad \forall i \in \{1, \dots, m\} \text{ and } \forall j \in \{1, \dots, n\} & (13)
\end{aligned}$$

We have

Proposition 3.1. *Let $\mathcal{H} = (V, \mathcal{E})$ be a hypergraph with edge weights $w : \mathcal{E} \rightarrow \mathbb{Q}_0^+$. The integer linear programs $\text{Min-}(b, n)\text{-ILP}$ and $\text{Min-}(b, k, \theta)\text{-ILP}$, $\theta > 0$, are equivalent to the minimum-weight perfect b -matching problem in \mathcal{H} .*

In the following we need some notations which we fix through the next remark.

Remark 3.2. Let $\text{Min-}(b, k, \theta)\text{-LP}$ be the linear programming relaxation of $\text{Min-}(b, k, \theta)\text{-ILP}$. Let $(b, k, \theta)\text{-ILP}$ be the system of inequalities built by the constraints (9) - (13) of $\text{Min-}(b, k, \theta)\text{-ILP}$, and let $(b, k, \theta)\text{-LP}$ be the LP-relaxation of $(b, k, \theta)\text{-ILP}$, where $X_i \in [0, 1] \cap \mathbb{Q}$ and $Y_j \in \mathbb{Q}_0^+$ for all i, j .

3.2 The Randomized Algorithm

Before we state the randomized algorithm, we have to ensure whether or not a $\text{Min-}(b, k, \theta)$ -matching exists. For a given b , a choice of $k = 0$ and $\theta = 0$ always makes the problem feasible. However, for some k and θ there might be no solution. Then we would like to find the maximum k such that a solution exists, given b and θ . Actually, for the integer programs we have to distinguish only between the cases $\theta = 0$ resp. $\theta > 0$ (which is the perfect b -matching problem).

ALGORITHM LP-SEARCH(θ)

Input: $\theta \geq 0$.

- 1) Test the solvability of $(b, 0, \theta)$ -LP. If it is not solvable, return " $(b, 0, \theta)$ -LP is not feasible." Otherwise set $k := 1$ and go to 2).
- 2) a) Test solvability of (b, k, θ) -LP and $(b, k + 1, \theta)$ -LP.
 b) If both are solvable, set $k := k + 2$ and go to 2a). If (b, k, θ) -LP is solvable, but $(b, k + 1, \theta)$ -LP is not solvable, return k . □

If $(b, 0, \theta)$ -LP is solvable, we define $k^* := \max \{k \in \mathbb{N}; k \leq n; (b, k, \theta)$ -LP has a solution $\}$. Obviously we have

Proposition 3.3. *The algorithm LP-SEARCH(θ) either outputs " $(b, 0, \theta)$ -LP is not feasible" or solving at most n LP's, it returns k^* .*

It is clear that the number of iterations can be dropped to at most $\lceil \log(n) \rceil$ using binary search.

In the following we work with a $k \in \mathbb{N}$, returned by the algorithm LP-search(θ), if it exists. The randomized rounding algorithm for the Min- (b, k, θ) -matching problem is the following:

ALGORITHM MIN- (b, k, θ) -RR

1. Solve the LP-relaxation Min- (b, k, θ) -ILP optimally, with solutions $x^* = (x_1^*, \dots, x_m^*)$ and $y^* = (y_1^*, \dots, y_n^*)$. Let $OPT^* = \sum_{i=1}^m w_i x_i^*$.
2. *Randomized Rounding:* Choose $\delta \in (0, 1]$. For $i = 1, \dots, m$, independently set the 0/1 random variable X_i to 1 with probability δx_i^* and to 0 with probability $1 - \delta x_i^*$. So $\Pr[X_i = 1] = \delta x_i^*$ and $\Pr[X_i = 0] = 1 - \delta x_i^*$, $\forall i \in \{1, \dots, m\}$.
3. Output X_1, \dots, X_m , the set of hyperedges $\mathcal{M}' = \{i \in \mathcal{E}; x_i = 1\}$, and its weight $w(\mathcal{M}')$. □

One can combine the algorithm Min- (b, k, θ) -RR with a greedy approach in order to get a feasible b -matching:

ALGORITHM MIN- (b, k, θ) -ROUND

- 1) Apply the algorithm MIN- (b, k, θ) -RR and output a set of hyperedges \mathcal{M}' .

- 2) List the nodes in a randomized order. Passing through this list and arriving at a node for which the b -matching condition is violated, we enforce the b -matching condition at this node by removing incident edges from \mathcal{M}' with highest weights.
- 3) Output is the so obtained set $\mathcal{M} \subseteq \mathcal{M}'$. □

Variants of this algorithm are possible, for example, one can remove edges incident in many nodes, etc.

4 Main Results and Proofs

4.1 The Main Results

We present an analysis of the algorithm MIN- (b, k, θ) -RR. Our most general result is the following theorem. c_1 and c_2 are positive constants depending only on l , δ , and θ . They will be specified more precisely later.

Theorem 4.1. *Let $\delta \in (0, 1)$ and $OPT^* \geq \frac{2}{3} \ln(4)(1 + 2\delta)(1 - \delta)^{-2}$. For $\lambda = \sqrt{\frac{m}{2} \ln(4n)}$ we have:*

- (a) *Let $\Delta \leq c_1 \cdot \frac{k}{b}$. For $\theta = 0$, the algorithm MIN- (b, k, θ) -RR returns a $(\delta b + \lambda)$ -matching \mathcal{M}' in \mathcal{H} of weight $w(\mathcal{M}') \leq OPT^*$ which covers at least*

$$\frac{k}{b}(1 - e^{-\delta b}) \left(1 - \sqrt{\frac{3b(\Delta(l-1) + 3)}{2k(1 - e^{-\delta b})}} \right) \quad (14)$$

nodes of \mathcal{H} with probability at least $1/4$.

- (b) *Let $\Delta \leq c_2 \cdot n$. For $\theta > 0$ the algorithm MIN- (b, k, θ) -RR returns a $(\delta b + \lambda)$ -matching \mathcal{M}' in \mathcal{H} of weight $w(\mathcal{M}') \leq OPT^*$ which covers at least*

$$0.632\delta\theta n \left(1 - \sqrt{\frac{2.38(\Delta(l-1) + 3)}{\delta\theta n}} \right) \quad (15)$$

nodes of \mathcal{H} with probability at least $1/4$.

For special b , we have a stronger result.

Theorem 4.2. *Let $\delta \in (0, 1)$. Assume that*

- i) $b \geq \frac{2}{3} \ln(4n)(1 + 2\delta)(1 - \delta)^{-2}$.
- ii) $OPT^* \geq \frac{2}{3} \ln(4)(1 + 2\delta)(1 - \delta)^{-2}$.

- (a) Let $\Delta \leq c_1 \cdot \frac{k}{b}$. For $\theta = 0$, the algorithm MIN- (b, k, θ) -RR returns a b -matching \mathcal{M}' in \mathcal{H} of weight $w(\mathcal{M}') \leq OPT^*$ which covers at least

$$\frac{k}{b}(1 - e^{-\delta b}) \left(1 - \sqrt{\frac{3b(\Delta(l-1) + 3)}{2k(1 - e^{-\delta b})}} \right)$$

nodes of \mathcal{H} with probability at least $1/4$.

- (b) Let $\Delta \leq c_2 \cdot n$. For $\theta > 0$ the algorithm MIN- (b, k, θ) -RR returns a b -matching \mathcal{M}' in \mathcal{H} of weight $w(\mathcal{M}') \leq OPT^*$ which covers at least

$$0.632\delta\theta n \left(1 - \sqrt{\frac{2.38(\Delta(l-1) + 3)}{\delta\theta n}} \right)$$

nodes of \mathcal{H} with probability at least $1/4$.

Remark 4.3. In Theorem 4.2 (a), for fixed δ , we have $b = \Omega(\ln(n))$. For $b = \Theta(\ln(n))$, and $k = \Omega(n)$ and $\Delta \leq c_1 \cdot \frac{k}{b}$ the number of covered nodes is at least

$$\Omega \left(\frac{n}{\ln(n)} (1 - o(1)) \right). \quad (16)$$

In this case we have an approximation of the maximum number of covered nodes k up to a factor of $1/\ln(n)$. From the techniques applied so far it is not clear whether the coverage can be improved towards $\Omega(k)$.

4.2 Proofs

We will first prove Theorem 4.2 and then 4.1. We start with a technical lemma.

Lemma 4.4. Let X_1, \dots, X_m be independent 0/1 random variables with $\mathbb{E}(X_i) = p_i \forall i = 1, \dots, m$. For $w_i \in [0, 1], i = 1, \dots, n, w(X) := \sum_{i=1}^m w_i X_i$. Let $z \geq 0$ be an upper bound on $\mathbb{E}(w(X))$, i.e., $\mathbb{E}(w(X)) \leq z$. Then

i) $\mathbb{P}[w(X) \geq z(1 + \beta)] \leq e^{-\frac{\beta^2 z}{2(1+\beta/3)}}$ for any $\beta > 0$.

ii) $\mathbb{P}[w(X) \geq z(1 + \beta)] \leq e^{-\frac{\beta^2 z}{3}}$ for $0 \leq \beta \leq 1$.

Proof: Let $z' := \lfloor z - \mathbb{E}(w(X)) \rfloor$, $p = z - \mathbb{E}(w(X)) - z'$, and let $Y_0, Y_1, \dots, Y_{z'}$ be independent 0/1 random variables with $\mathbb{E}(Y_0) = p$ and $Y_j = 1 \quad \forall j \geq 1$. The random variable $X' := w(X) + Y_0 + Y_1 + \dots + Y_{z'}$ satisfies $\mathbb{E}(X') = z$ and $X' \geq w(X)$ and we may apply the Angluin-Valiant inequality (Theorem 2.8) to it:

i) For any $\beta > 0$ we have

$$\mathbb{P}[w(X) \geq z(1 + \beta)] \leq \mathbb{P}[X' \geq z(1 + \beta)] \leq e^{-\frac{\beta^2 z}{2(1+\beta/3)}}.$$

ii) For $0 \leq \beta \leq 1$ it is easy to see that $e^{-\frac{\beta^2 z}{2(1+\beta/3)}} \leq e^{-\frac{\beta^2 z}{3}}$. \square

Let X_1, \dots, X_m and \mathcal{M}' be the output of the algorithm MIN- (b, k, θ) -RR. Further let OPT and OPT^* be the integer resp. LP-optima for Min- (b, k, θ) -ILP.

Lemma 4.5. *Suppose that $\delta \in (0, 1)$ and $b \geq \frac{2}{3} \ln(4n)(1 + 2\delta)(1 - \delta)^{-2}$. Then*

$$\mathbb{P} \left[\exists j \in V : \sum_{i=1}^m a_{ij} X_i \geq b \right] \leq \frac{1}{4}.$$

Proof: First we compute the expectation.

$$\mathbb{E} \left(\sum_{i=1}^m a_{ij} X_i \right) = \sum_{i=1}^m a_{ij} \mathbb{E}(X_i) = \sum_{i=1}^m a_{ij} \delta x_i^* = \delta \cdot \sum_{i=1}^m a_{ij} x_i^* \leq \delta b. \quad (17)$$

Set $\beta := \frac{1}{\delta} - 1$. With Lemma 4.4 we get:

$$\begin{aligned} \mathbb{P} \left[\sum_{i=1}^m a_{ij} X_i \geq b \right] &= \mathbb{P} \left[\sum_{i=1}^m a_{ij} X_i \geq (1 + \beta) \delta b \right] \\ &\leq \exp \left(\frac{-\beta^2 \delta b}{2(1 + \beta/3)} \right) \\ &= \exp \left(-\frac{3}{2} \cdot \frac{(1 - \delta)^2}{1 + 2\delta} \cdot b \right) \\ &\leq \frac{1}{4n} \quad (\text{using the assumption on } b) \end{aligned}$$

So

$$\mathbb{P} \left[\exists j \in V : \sum_{i=1}^m a_{ij} X_i \geq b \right] \leq \sum_{j=1}^n \mathbb{P} \left[\sum_{i=1}^m a_{ij} X_i \geq b \right] \leq n \cdot \left(\frac{1}{4n} \right) = \frac{1}{4}. \quad \square$$

Lemma 4.6. *Suppose that $\delta \in (0, 1)$ and $OPT^* \geq \frac{2}{3} \ln(4)(1 + 2\delta)(1 - \delta)^{-2}$. Then*

$$\mathbb{P} \left[\sum_{i=1}^m w_i X_i \geq OPT^* \right] \leq \frac{1}{4}.$$

Proof: We have

$$\mathbb{E} \left(\sum_{i=1}^m w_i X_i \right) = \sum_{i=1}^m w_i \mathbb{E}(X_i) = \sum_{i=1}^m w_i \delta x_i^* = \delta \cdot \sum_{i=1}^m w_i x_i^* = \delta \cdot OPT^*. \quad (18)$$

Choose $\beta = \frac{1}{\delta} - 1$. Then

$$\begin{aligned} \mathbb{P} \left[\sum_{i=1}^m w_i X_i \geq OPT^* \right] &= \mathbb{P} \left[\sum_{i=1}^m w_i X_i \geq \delta(1 + \beta) OPT^* \right] \\ &= \mathbb{P} \left[\sum_{i=1}^m w_i X_i \geq \mathbb{E} \left(\sum_{i=1}^m w_i X_i \right) (1 + \beta) \right] \\ &\leq \exp \left(\frac{-\beta^2 \mathbb{E}(\sum_{i=1}^m w_i X_i)}{2(1 + \beta/3)} \right) \text{ (Theorem 2.8(i))} \\ &= \exp \left(\frac{-\beta^2 \delta OPT^*}{2(1 + \beta/3)} \right) \\ &\leq \frac{1}{4} \end{aligned}$$

where the last inequality follows from the assumption on OPT^* . \square

We now come to a key lemma, which controls the covering quality of the randomized algorithm.

Let $Y_j := \sum_{i=1}^m a_{ij} X_i$ for all j and $Y := \sum_{j=1}^n Y_j$.

Lemma 4.7. *For any $\delta \in (0, 1]$,*

- i) $\mathbb{E}(Y) \geq n - \sum_{j=1}^n e^{-\delta \sum_{i=1}^m a_{ij} x_i^*}$
- ii) *If $\theta > 0$, then $\mathbb{E}(Y) \geq n(1 - e^{-\delta\theta}) \geq 0.632\delta\theta n$.*
- iii) *For Min- (b, k, θ) -RR with $\theta = 0$ we have $\mathbb{E}(Y) \geq \frac{k}{b}(1 - e^{-\delta b})$.*

Proof:

- i) Define $\mathcal{E}_j := \{E \in \mathcal{E}; j \in E\}$. We have

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E} \left(\sum_{j=1}^n Y_j \right) = \sum_{j=1}^n \mathbb{E}(Y_j) = \sum_{j=1}^n \mathbb{P}[Y_j = 1] \\ &= \sum_{j=1}^n (1 - \mathbb{P}[Y_j = 0]) = n - \sum_{j=1}^n \mathbb{P}[Y_j = 0] \end{aligned} \quad (19)$$

Now

$$\begin{aligned}
\mathbb{P}[Y_j = 0] &= \mathbb{P}\left[\sum_{i=1}^m a_{ij}X_i = 0\right] \\
&= \mathbb{P}[(a_{1j}X_1 = 0) \wedge \dots \wedge (a_{mj}X_m = 0)] \\
&= \prod_{i=1}^m \mathbb{P}[a_{ij}X_i = 0] \\
&= \prod_{i \in \mathcal{E}_j} \mathbb{P}[X_i = 0] = \prod_{i \in \mathcal{E}_j} (1 - \delta x_i^*).
\end{aligned}$$

For $u \in \mathbb{R}$ we have the inequality $1 - u \leq e^{-u}$. Thus

$$\prod_{i \in \mathcal{E}_j} (1 - \delta x_i^*) \leq \prod_{i \in \mathcal{E}_j} e^{-\delta x_i^*} = e^{-\delta \cdot \sum_{i=1}^m a_{ij}x_i^*} \quad (20)$$

Hence, with (19),(20) and (20)

$$\mathbb{E}(Y) \geq n - \sum_{j=1}^n e^{-\delta \cdot \sum_{i=1}^m a_{ij}x_i^*} \quad (21)$$

- ii) Since $\sum_{i=1}^m a_{ij}x_i^* \geq \theta$ for all $j \in \{1, \dots, n\}$, the first inequality immediately follows from (i). For the second inequality, observe that for $x \in [0, 1]$, $e^{-x} \leq 1 - x + x/e$. This is true, because the linear function $1 - x + x/e$ is an upper bound for the convex function e^{-x} in $[0, 1]$. So $\mathbb{E}(Y) \geq (1 - e^{-\delta\theta})n \geq (1 - 1/e)\delta\theta n \geq 0.632\delta\theta n$.
- iii) Since e^{-x} is convex, the linear function $1 - x(\delta b)^{-1} + xe^{-\delta b}(\delta b)^{-1}$ is an upper bound for e^{-x} in $[0, \delta b]$. With (21) we get

$$\mathbb{E}(Y) \geq n - n + (1 - e^{-\delta b})\delta(\delta b)^{-1} \underbrace{\sum_{j=1}^n \sum_{i=1}^m a_{ij}x_i^*}_{\geq k} \geq (1 - e^{-\delta b}) \cdot \frac{k}{b}. \quad \square$$

An upper bound for the variance of Y can be computed directly, via co-variance and dependent pairs:

Lemma 4.8. *Let Δ be the maximum vertex degree of \mathcal{H} and let l be the maximum cardinality of a hyperedge. Then*

$$\text{Var}(Y) \leq \frac{1}{2} \cdot (\Delta(l - 1) + 3)\mathbb{E}(Y)$$

Proof: By Proposition 2.6

$$\text{Var}(Y) \leq \mathbb{E}(Y) + 2\gamma \quad (22)$$

where γ is the sum of $\mathbb{E}(Y_i Y_j)$ of all unordered dependent pairs i, j . Since the Y_i 's are 0/1 random variables, we have for pairs $\{i, j\}$ with $i \sim j$

$$\begin{aligned} \mathbb{E}(Y_i Y_j) &= \mathbb{P}[(Y_i = 1) \wedge (Y_j = 1)] \\ &\leq \min(\mathbb{P}[Y_i = 1], \mathbb{P}[Y_j = 1]) \\ &\leq \frac{1}{2}(\mathbb{P}[Y_i = 1] + \mathbb{P}[Y_j = 1]) \\ &= \frac{1}{2}(\mathbb{E}(Y_i) + \mathbb{E}(Y_j)) \end{aligned}$$

Hence

$$\begin{aligned} \gamma &= \sum_{\{i,j\} \in \Gamma} \mathbb{E}(Y_i Y_j) \leq \frac{1}{2} \sum_{\{i,j\} \in \Gamma} (\mathbb{E}(Y_i) + \mathbb{E}(Y_j)) \\ &\leq \frac{1}{4} \sum_{i=1}^n \left(\mathbb{E}(Y_i) + \sum_{j \sim i} \mathbb{E}(Y_j) \right) = \frac{1}{4} \mathbb{E}(Y) + \frac{1}{4} \sum_{i=1}^n \sum_{j \sim i} \mathbb{E}(Y_j) \\ &\leq \frac{1}{4} \mathbb{E}(Y) + \frac{1}{4} \sum_{i=1}^n \mathbb{E}(Y_i) \Delta(l-1) = \frac{1}{4} \mathbb{E}(Y) + \frac{1}{4} \Delta(l-1) \mathbb{E}(Y) \\ &= \frac{1}{4} (\Delta(l-1) + 1) \mathbb{E}(Y), \end{aligned}$$

and (22) concludes the proof. \square

Let c_1 and c_2 be positive constants depending only on l, δ , and θ such that for $\Delta \leq c_1 \cdot \frac{k}{b}$ resp. $\Delta \leq c_2 \cdot n$ we have

$$\sqrt{\frac{3b(\Delta(l-1) + 3)}{2k(1 - e^{-\delta b})}} < 1 \text{ resp. } \sqrt{\frac{2.38(\Delta(l-1) + 3)}{\delta \theta n}} < 1. \quad (23)$$

Note that in the following we assume l, δ , and θ to be constants.

Lemma 4.9.

i) For $a := \sqrt{\frac{3}{2}(\Delta(l-1) + 3)\mathbb{E}(Y)}$: $\mathbb{P}[Y \leq \mathbb{E}(Y) - a] \leq \frac{1}{4}$.

ii) Let $\Delta \leq c_1 \cdot \frac{k}{b}$. Then

$$\mathbb{E}(Y) - a \geq \frac{k}{b}(1 - e^{-\delta b}) \left(1 - \sqrt{\frac{3b(\Delta(l-1) + 3)}{2k(1 - e^{-\delta b})}} \right) \text{ if } \theta = 0.$$

iii) Let $\Delta \leq c_2 \cdot n$. Then

$$\mathbb{E}(Y) - a \geq 0.632\delta\theta n \left(1 - \sqrt{\frac{2.38(\Delta(l-1)+3)}{\delta\theta n}} \right) \quad \text{if } \theta > 0.$$

Proof: i) With the Chebyshev-Cantelli inequality (Theorem 2.5) we have

$$\begin{aligned} \mathbb{P}[Y \leq \mathbb{E}(Y) - a] &\leq \frac{\text{Var}(Y)}{\text{Var}(Y) + a^2} = \frac{1}{1 + \frac{a^2}{\text{Var}(Y)}} \\ &\leq \frac{1}{1 + \frac{a^2}{0.5(\Delta(l-1)+3)\mathbb{E}(Y)}} \quad (\text{Lemma 4.8}) \\ &= \frac{1}{1 + \frac{1.5(\Delta(l-1)+3)\mathbb{E}(Y)}{0.5(\Delta(l-1)+3)\mathbb{E}(Y)}} = \frac{1}{4}. \end{aligned}$$

ii) and iii):

$$\mathbb{E}(Y) - a = \mathbb{E}(Y) \left(1 - \frac{\sqrt{\frac{3}{2}(\Delta(l-1)+3)\mathbb{E}(Y)}}{\mathbb{E}(Y)} \right) \quad (24)$$

$$= \mathbb{E}(Y) \left(1 - \frac{\sqrt{3(\Delta(l-1)+3)}}{\sqrt{2}\mathbb{E}(Y)} \right) \quad (25)$$

$$\geq \begin{cases} \frac{k}{b}(1 - e^{-\delta b}) \left(1 - \sqrt{\frac{3b(\Delta(l-1)+3)}{2k(1 - e^{-\delta b})}} \right) & \text{if } \theta = 0 \\ 0.632\delta\theta n \left(1 - \sqrt{\frac{2.38(\Delta(l-1)+3)}{\delta\theta n}} \right) & \text{if } \theta > 0 \end{cases} \quad (26)$$

Note that to the upper bound condition for Δ , the lower bounds in (26) are positive. \square

Proof of Theorem 4.2: Lemma 4.5, 4.6 and Lemma 4.9 imply the theorem. \square

This theorem holds only for $b = \Omega(\ln(n))$. In the rest of this section we give an analysis also for the case of arbitrary b , losing a certain amount of feasibility.

Lemma 4.10. Let $\delta > 0$, $\mu_j = \mathbb{E}(\sum_{i=1}^m a_{ij}X_i)$ for all j , and $\lambda = \sqrt{\frac{m}{2} \ln(4n)}$. Then

$$\mathbb{P} \left[\exists j : \sum_{i=1}^m a_{ij}X_i > \delta b + \lambda \right] \leq \frac{1}{4}.$$

Proof: As in (17), $\mu_j = \mathbb{E}(\sum_{i=1}^m a_{ij}X_i) \leq \delta b$ for all j . With the Chernoff-Hoeffding bound (Theorem 2.7)

$$\mathbb{P} \left[\sum_{i=1}^m a_{ij}X_i > \delta b + \lambda \right] \leq \mathbb{P} \left[\sum_{i=1}^m a_{ij}X_i > \mu_j + \lambda \right] \leq \exp \left(\frac{-2\lambda^2}{m} \right) = \exp(-\ln(4n)) = \frac{1}{4n}.$$

So,

$$\mathbb{P} \left[\exists j \in V : \sum_{i=1}^m a_{ij} X_i > \delta b + \lambda \right] \leq \sum_{j=1}^n \mathbb{P} \left[\sum_{i=1}^m a_{ij} X_i > \delta b + \lambda \right] \leq n \cdot \left(\frac{1}{4n} \right) = \frac{1}{4}. \quad \square$$

Proof of Theorem 4.1: Lemma 4.6, 4.9 and Lemma 4.10 imply the theorem. \square

5 Experimental Results

5.1 Implementation

We run the algorithm $\text{MIN-}(b, k, \theta)\text{-ROUND}$ for different values of b , k , and θ in a C^{++} -implementation.

Recall that the algorithm $\text{MIN-}(b, k, \theta)\text{-ROUND}$ has 3 steps:

1. It solves the linear program and delivers a fractional solution,
2. applies randomized rounding on the fractional solution and delivers an integer solution, which we call *primary solution*,
3. removes edges from the primary solution.

In the primary solution the nodes might be covered by more than b edges. The superfluous edges are removed in step 3. Edges are removed in a randomized greedy approach. The nodes are chosen in a randomized order and if the considered node is covered by more than one edge, the ones with the greatest cost are removed. In the following tables we use 100 runs of the randomized rounding algorithm. As the final solution, we choose the one with the fewest number of uncovered nodes. (If this choice is not unique, we pick one with the smallest cost.) The LP's are solved with a simplex-method with the CLP-solver from the free COIN-OR library [1].

The columns in the tables in Sections 5.2 and 5.3 are organized as follows:

- 1: represents patient data
(Patient 1-5 are real patient data, whereas patient 6 is a phantom.)
- 2: represents number of seeds to be matched
- 3: represents the cost of the LP-solution
- 4: represents the cost of the matching returned by the algorithm
- 5: represents the running time in CPU seconds of the program
- 6: represents number of unmatched seeds

5.2 Results for the Algorithm Min- $(b, k, 0)$ -Round

1. $b = 1.20, k = 0.90 \cdot n, \theta = 0.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	54.31	48.30	14.23	12
2	67	77.03	66.02	14.61	14
3	31	17.08	16.62	1.01	4
4	22	17.96	17.18	0.32	3
5	43	72.39	53.43	3.40	13
6	25	5.30	4.42	0.50	6

2. $b = 1.20, k = 1.00 \cdot n, \theta = 0.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	67.77	58.20	14.73	7
2	67	93.32	78.22	14.54	9
3	31	21.18	19.44	1.02	2
4	22	21.94	19.24	0.33	2
5	43	91.34	55.33	3.59	14
6	25	6.51	5.44	0.51	4

3. $b = 1.10, k = 0.90 \cdot n, \theta = 0.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	56.40	54.11	14.23	9
2	67	80.20	75.61	14.60	10
3	31	17.41	16.62	1.00	4
4	22	18.20	17.18	0.32	3
5	43	79.77	63.41	3.61	11
6	25	5.71	4.91	0.52	5

4. $b = 1.10, k = 1.00 \cdot n, \theta = 0.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	71.59	63.12	15.23	5
2	67	98.13	88.89	15.52	5
3	31	21.84	20.97	1.01	1
4	22	22.78	21.50	0.34	1
5	43	103.85	68.53	4.10	12
6	25	7.22	6.69	0.51	2

5. $b = 1.00, k = 0.90 \cdot n, \theta = 0.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	58.91	60.56	14.81	6
2	67	84.17	86.11	14.34	6
3	31	17.80	17.94	1.01	3
4	22	18.83	19.24	0.34	2
5	43	91.67	66.53	3.95	13
6	25	6.37	6.69	0.51	2

6. $b = 1.00$, $k = 1.00 \cdot n$, $\theta = 0.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	80.60	80.60	15.59	0
2	67	170.27	119.53	33.48	8
3	31	22.58	22.58	1.01	0
4	22	23.82	23.82	0.33	0
5	43	199.65	199.65	7.40	0
6	25	8.85	8.85	0.54	0

5.3 Results for the Algorithm Min- (b, k, θ) -Round with $\theta > 0$

7. $b = 1.20$, $k = 0.90 \cdot n$, $\theta = 0.20$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	55.27	55.10	14.26	9
2	67	80.33	74.11	14.34	11
3	31	17.29	18.57	1.01	3
4	22	18.17	19.24	0.33	2
5	43	74.77	51.90	3.49	14
6	25	5.51	3.99	0.52	7

8. $b = 1.20$, $k = 1.00 \cdot n$, $\theta = 0.20$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	68.25	60.76	14.31	6
2	67	96.20	78.11	14.73	9
3	31	21.22	20.97	1.01	1
4	22	21.94	19.24	0.33	2
5	43	92.77	55.19	3.70	13
6	25	6.66	4.91	0.52	5

9. $b = 1.10$, $k = 0.90 \cdot n$, $\theta = 0.20$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	57.41	57.01	14.17	8
2	67	83.51	75.54	14.48	11
3	31	17.61	19.73	1.01	2
4	22	18.45	19.24	0.32	2
5	43	81.39	55.20	3.64	13
6	25	5.89	4.48	0.51	6

10. $b = 1.10$, $k = 1.00 \cdot n$, $\theta = 0.20$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	71.79	63.12	14.88	5
2	67	100.78	93.04	15.32	4
3	31	21.85	22.58	1.02	0
4	22	22.79	23.82	0.33	0
5	43	104.57	72.66	3.89	13
6	25	7.32	6.69	0.52	2

11. $b = 1.00$, $k = 0.90 \cdot n$, $\theta = 0.20$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	59.93	63.64	14.19	5
2	67	87.51	87.97	14.43	6
3	31	17.94	20.97	1.02	1
4	22	18.92	21.56	0.33	1
5	43	92.40	58.17	3.91	15
6	25	6.48	6.69	0.52	2

12. $b = 1.00$, $k = 1.00 \cdot n$, $\theta = 0.20$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	80.60	80.60	15.92	0
2	67	170.27	115.10	31.82	8
3	31	22.58	22.58	1.01	0
4	22	23.82	23.82	0.33	0
5	43	199.65	199.65	7.77	0
6	25	8.85	8.85	0.52	0

13. $b = 1.20$, $k = 0.90 \cdot n$, $\theta = 1.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	79.80	76.20	14.49	1
2	67	122.16	111.16	15.96	2
3	31	22.58	22.58	1.01	0
4	22	23.78	23.82	0.33	0
5	43	129.47	77.03	4.58	12
6	25	8.80	8.85	0.52	0

14. $b = 1.20, k = 1.00 \cdot n, \theta = 1.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	79.80	76.20	14.83	1
2	67	122.16	111.16	16.34	2
3	31	22.58	22.58	1.04	0
4	22	23.78	23.82	0.34	0
5	43	129.47	77.62	4.54	12
6	25	8.80	8.85	0.53	0

15. $b = 1.10, k = 0.90 \cdot n, \theta = 1.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	80.20	80.60	14.49	0
2	67	125.93	108.72	16.90	3
3	31	22.58	22.58	1.01	0
4	22	23.80	23.82	0.34	0
5	43	142.54	92.19	5.09	11
6	25	8.83	8.85	0.50	0

16. $b = 1.10, k = 1.00 \cdot n, \theta = 1.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	80.20	80.60	15.37	0
2	67	125.93	108.72	17.42	3
3	31	22.58	22.58	1.04	0
4	22	23.80	23.82	0.33	0
5	43	142.54	90.05	4.98	11
6	25	8.83	8.85	0.52	0

17. $b = 1.00, k = 0.90 \cdot n, \theta = 1.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	80.60	80.60	14.74	0
2	67	170.27	111.85	25.48	8
3	31	22.58	22.58	1.01	0
4	22	23.82	23.82	0.33	0
5	43	199.65	199.65	6.50	0
6	25	8.85	8.85	0.52	0

18. $b = 1.00$, $k = 1.00 \cdot n$, $\theta = 1.00$

Patient	Seeds	LP-OPT	Cost	Time	Unmatched
1	67	80.60	80.60	14.24	0
2	67	170.27	113.91	25.65	7
3	31	22.58	22.58	0.99	0
4	22	23.82	23.82	0.32	0
5	43	199.65	199.65	6.48	0
6	25	8.85	8.85	0.51	0

5.4 Discussion – Implementation vs. Theory

With the algorithm $\text{MIN-}(b, k, \theta)\text{-ROUND}$ for $\theta = 0$ we get optimal results (except patient 2) if the constraints are most restrictive: $b = 1.00$, $k = 1.00 \cdot n$, see Table 6. With the algorithm $\text{MIN-}(b, k, \theta)\text{-ROUND}$ for $\theta > 0$, the same observation holds: optimal results (except patient 2) are achieved with the most restrictive constraints: $b = 1.00$, $k = 1.00 \cdot n$, $\theta = 0.2$ (Table 12) and $b = 1.00$, $k = 1.00 \cdot n$, $\theta = 1.00$ (Table 18). Obviously, a high θ can compensate for a low k , and vice versa (see Table 5 in comparison to 17, and 5 in comparison to 6).

This clearly shows that the practical results for the instances are much better than the analysis of Section 4 indicates. However, to close the gap between theory and practice seems to be a challenging problem in the area of randomized algorithms, where the so far developed probabilistic tools seem to be insufficient.

The non-optimal results for patient 2 could be explained by the bad image quality of the X-rays and movement of the patient in the time between taking two different X-rays. Since it is important to find the correct matching of the seeds and not just any minimum-weight perfect matching, the question if this is really the right matching is legitimate. This is difficult to prove, but results with help of a graphical 3D-program seem to be promising: we take the proposed seed positions in 3D and produce pictures, showing how the seeds would lie on the X-rays, if these were the real positions. A comparison between the pictures and the real X-rays shows that the positions agree.

This observation is supported by the results for the phantom (patient 6), where we know the seed positions, and where the algorithm returns the optimal solution, see, e.g., Table 18.

The running times of our algorithm are of the same order of magnitude as those of the commercial software VariSeed[®] presently used at the Clinic of Radiotherapy. These range between 4 and 20 seconds for instances with 43 to 82 seeds resp. However – due to technical and licensing issues – we had to measure these times on a different computer (different CPU and operating system, but approximately the same frequency) than the one where the tests for our algorithm were performed, and we also had no exact method of measurement available (just a stopwatch). As we are dealing with an offline application, a few seconds in running time are unimportant. Also our implementation can likely be improved (especially the part for reading in the large instance files) to even gain a few seconds in running time and possibly clearly outperform the commercial software with respect to running time.

Our main advantage, however, lies in the quality of the solution delivered. Our algorithm delivered the correct solution also in certain cases where the commercial one failed. As shown by Siebert et al. [24], VariSeed[®] (versions 6.7/7.1) can compute wrong 3D seed distributions if seeds are arranged in certain ways, and these errors cannot be explained by the ambiguities inherent to the three-film technique. Our algorithm however performs well on the phantom instance studied in [24] (as well as on the tested patient data, except for patient 2, which had a poor image quality). As a consequence, the immigration of our algorithm in the brachytherapy planning process at the Clinic of Radiotherapy in Kiel is planned.

6 Open Problems

Most interesting are the following problems, which we leave open but would like to discuss in future work.

1. At the moment, we can analyze the randomized rounding algorithm, but we are not able to analyze the repairing step of the algorithm MIN- (b, k, θ) -ROUND. But this of course is a major challenge for future work.
2. Can the coverage of $\Omega\left(\frac{k}{b}\right)$ in Theorem 4.2 be improved towards $\Omega(k)$?
3. Can the b -matching lower bound assumption $b = \Omega(\ln(n))$ in Theorem 4.2 be dropped towards $b = O(1)$?
4. What is the approximation complexity of the minimum-weight perfect matching problem in hypergraphs? Is there a complexity-theoretic threshold?

References

- [1] COIN-OR library, <http://www.coin-or.org>
- [2] N. Alon, J. Spencer, P. Erdős; *The probabilistic method*. John Wiley & Sons, Inc., 1992

- [3] M.D. Altschuler, R.D. Epperson, P.A. Findlay; *Rapid, accurate, three-dimensional location of multiple seeds in implant radiotherapy treatment planning*. Phys. Med. Biol. 28, 1305-1318, 1983
- [4] H.I. Amols, I.I. Rosen; *A three-film technique for reconstruction of radioactive seed implants*. Med.Phys 8(2), 1981
- [5] H.I. Amols, G.N. Cohen, D.A. Todor, M. Zaider; *Operator-free, film-based 3D seed reconstruction in brachytherapy* Phys. Med. Biol. 47(12), 2031-2048, 2002
- [6] D. Angluin, L.G. Valiant; *Fast probabilistic algorithms for Hamiltonian circuits and matchings*. J. Computer System Sciences 18, 155-193, 1979
- [7] D. Ash, J. Battermann, L. Blank, A. Flynn, T. de Reijke, P. Lavagnini; *ES-TRO/EAU/EORTC recommendations on permanent seed implantation for localized prostate cancer*. Radiother Oncol 57, 315-321, 2000
- [8] E. Balas, M.J. Saltzman; *An algorithm for the three-index assignment problem*. Operations Research 39(1), 1991
- [9] L. Beaulieu, J. Pouliot, D. Tubic, A. Zaccarin; *Automated seed detection and three-dimensional reconstruction. II. Reconstruction of permanent prostate implants using simulated annealing*. Phys. Med. 28, 2272-2279
- [10] P.J. Biggs, D.M. Kelley; *Geometric reconstruction of seed implants using a three-film technique*. Med. Phys. 10, 701-705, 1983
- [11] W.L. Brogan; *Algorithm for Ranked Assignments with Applications to Multiobject Tracking*. J. Guidance 12(3) 1989
- [12] H. Chernoff; *A measure of asymptotic efficiency for test of a hypothesis based on the sum of observation*. Ann. Math. Stat. 23, 493-509, 1952
- [13] L.M. Chin, R.L. Siddon; *Two-film brachytherapy reconstruction algorithm*. Med. Phys. 12, 77-83, 1985
- [14] P.S. Cho, R.J. Marks, S. Narayanan; *Three-dimensional seed reconstruction from an incomplete data set for prostate brachytherapy*. Phys. Med. Biol. 49, 3483-3494, 2004
- [15] P.S. Cho, S.T. Lam, R.J. Marks II, S. Narayanan; *3D Seed reconstruction for prostate brachytherapy using hough trajectories*. Phys. Med. Biol 49, 557-569, 2004
- [16] M.R. Garey, D.S. Johnson; *Computers and Intractability*. W.H. Freeman and Company, New York, 1979
- [17] M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, B. Reed; *Probabilistic methods for algorithmic discrete mathematics*. Springer Series Algorithms and Combinatorics, 16, Springer Verlag, 1998

- [18] W. Hoeffding; *Probability inequalities for sums of bounded random variables*. J. Amer. Statist. Assoc. 58, 13-30, 1963
- [19] S. Janson, T. Luczak, A. Ruciński; *Random Graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, Inc. New York Toronto 2000
- [20] E.K. Lee, R.J. Gallagher, D. Silvern, C.S. Wu, M. Zaider; *Treatment planning for brachytherapy: an integer programming model, two computational approaches and experiments with permanent prostate implant planning*. Physics in Medicine and Biology, 44 (1), 145-165, 1999
- [21] R. Nath, M.S. Rosenthal; *An automatic seed identification technique for interstitial implants using three isocentric radiographs*. Med. Phys. 10(4), 1983
- [22] M. Okamoto. *Some inequalities relating to the partial sum of binomial probabilities*. Ann. Inst. Statist. Math. 10, 29-35, 1958
- [23] P. Raghavan, C.D. Thompson; *Randomized rounding: a technique for provably good algorithms and algorithmic proofs*. Combinatorica 7 (4), 365-374, 1987
- [24] F.-A. Siebert, P. Kohr, G. Kovács; *The design and testing of a solid phantom for the verification of a commercial 3D seed reconstruction algorithm*. Radiotherapy and Oncology 74, 169-175, 2005

