Open Problems and Directions in the Analysis of Partial Match

Amalia Duch, Gustavo Lau and Conrado Martínez

Department of Computer Science
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain

AofA 2017
Princeton, USA

Introduction

- Our history of PM queries over hierarchical multidimensional data structures (DS) starts in 1974 with Finkel's PhD dissertation and the introduction of quadtrees by Bentley and Finkel.

R. A. Finkel

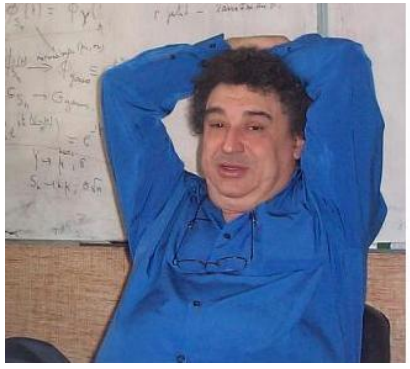
J. L. Bentley

- In 1975 Bentley introduced the K-d trees.

$$\text{Expected cost PM} \sim \beta n^{1-s/K}$$

Introduction

- The seminal paper on the study of the expected cost of PM queries started in 1986 with the seminal paper of Flajolet and Puech in K-d trees and K-d tries.



Ph. Flajolet

C. Puech

$$\text{Expected cost PM} \sim \beta n^{1-s/K+\theta(s/K)} = \beta n^{\alpha}$$

PM queries: values, patterns and ranks

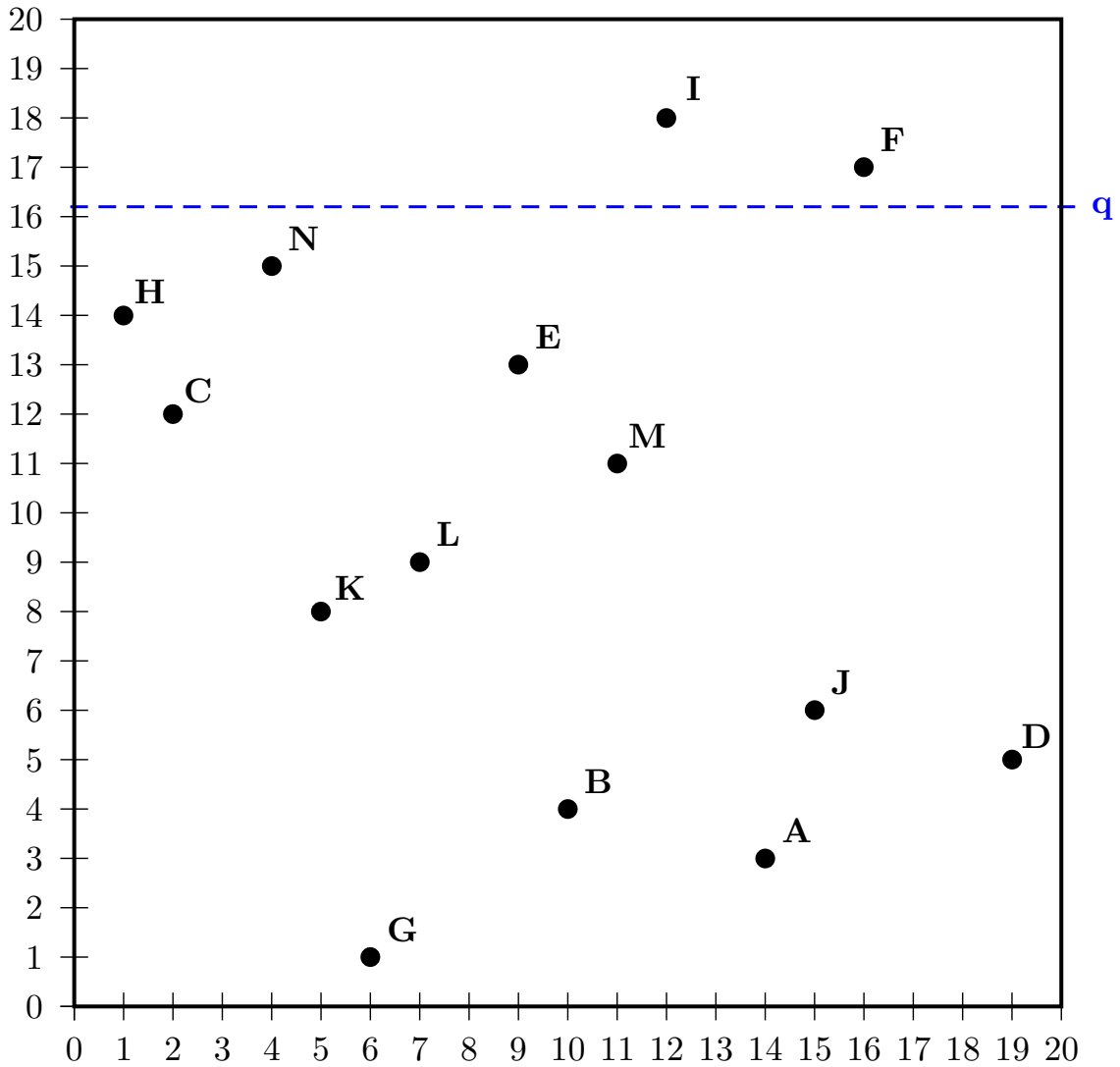
Given a file $F \subset \mathcal{D}_0 \times \cdots \times \mathcal{D}_{K-1}$:

A **partial match** query $\mathbf{q}$ is given by $\mathbf{q} = (q_0, \ldots, q_{K-1})$ with $q_i \in \mathcal{D}_i \cup \{*\}$. The coordinates $q_i \neq *$ are called **specified**, otherwise they are called **unspecified**. Assumption: the number s of specified coordinates satisfies $0 < s < K$.

The **query pattern** $\mathbf{u}(\mathbf{q}) = (u_0, \ldots, u_{K-1})$ of $\mathbf{q}$ is such that $u_i = S$ if $q_i \neq *$ and $u_i = *$ otherwise.

The **rank vector** $\mathbf{q}$ is the vector $\mathbf{r}(F, \mathbf{q}) = \mathbf{r}(\mathbf{q}) = (r_0, \ldots, r_{K-1})$ where $r_i = *$ when $q_i = *$ and r_i is the number of records $\mathbf{x}$ in F such that $x_i \leq q_i$ when $q_i \neq *$.

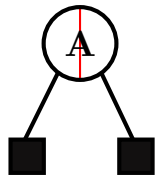
PM query example

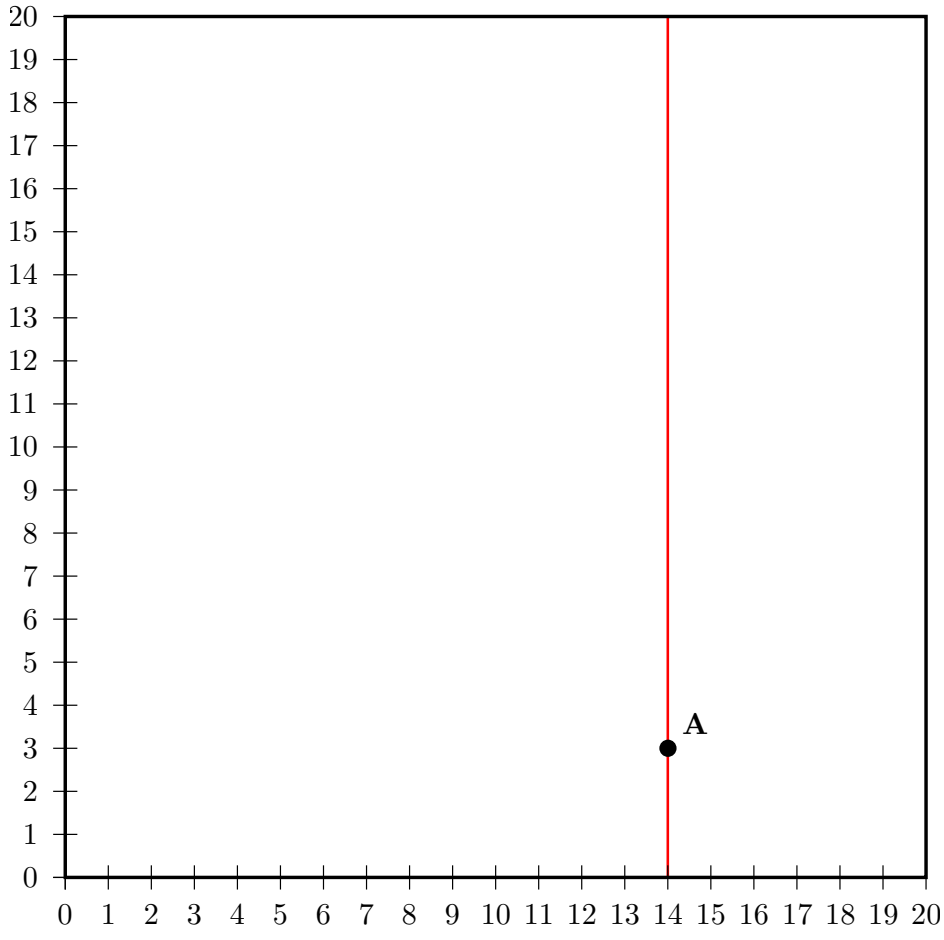


$K = 2, n = 14$

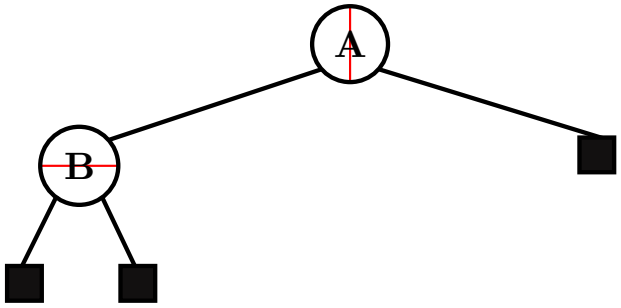
Query: $\mathbf{q} = (*, 16.2)$, Pattern: $\mathbf{u} = (*, S)$, Rank: $\mathbf{r} = (*, 12)$

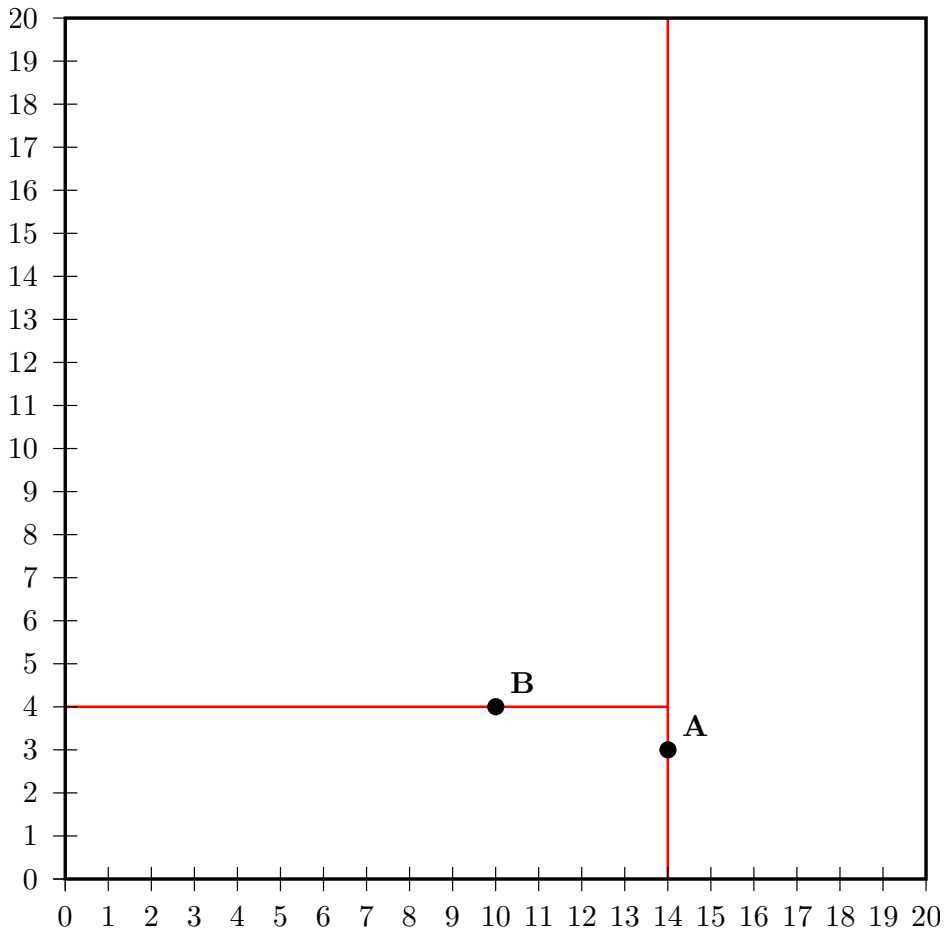
Example: K-d tree and induced partition



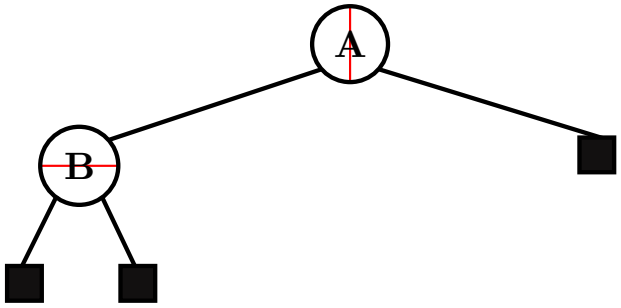


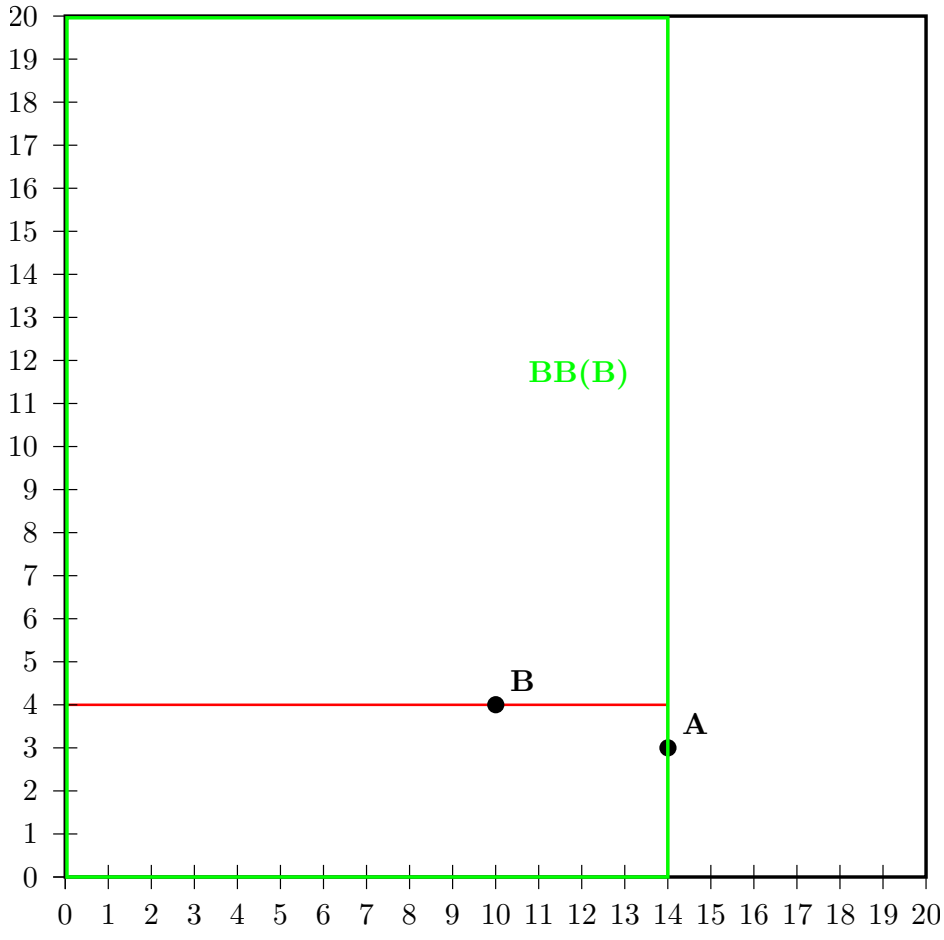
Example: K-d tree and induced partition



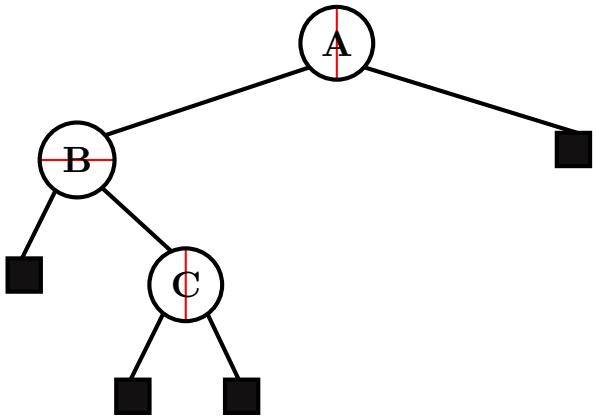


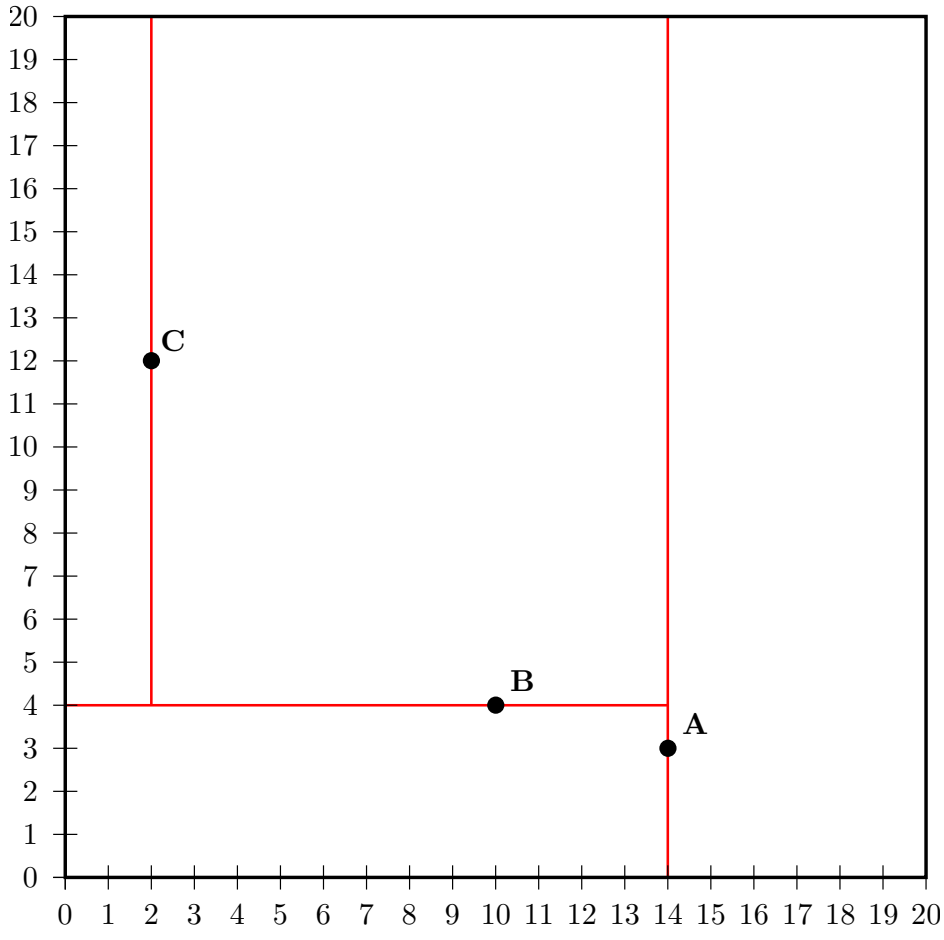
Example: K-d tree and induced partition



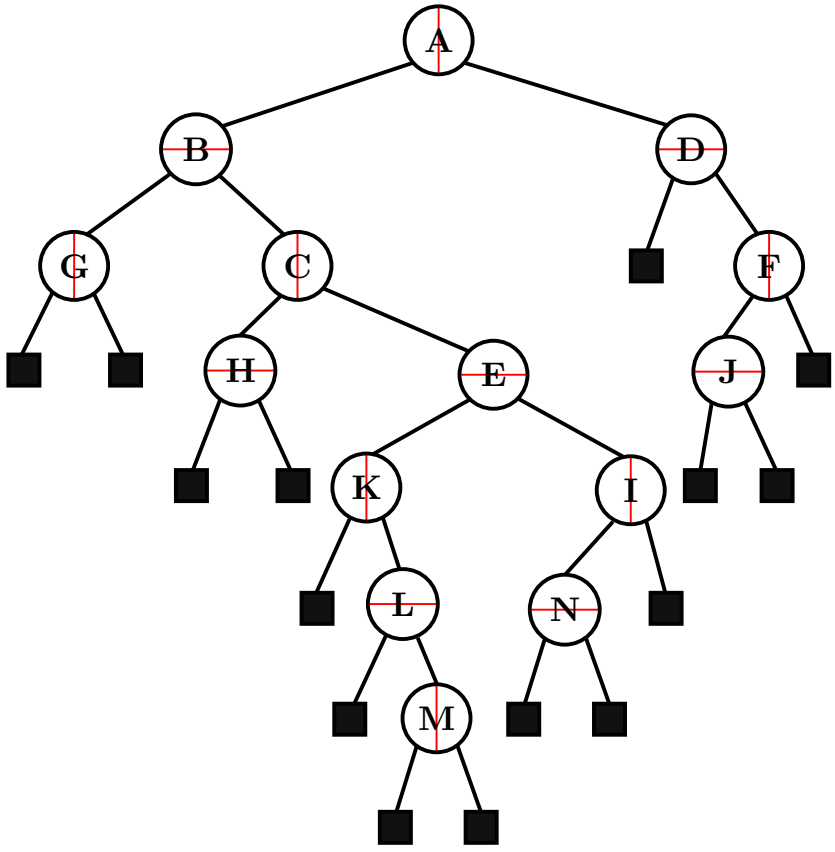


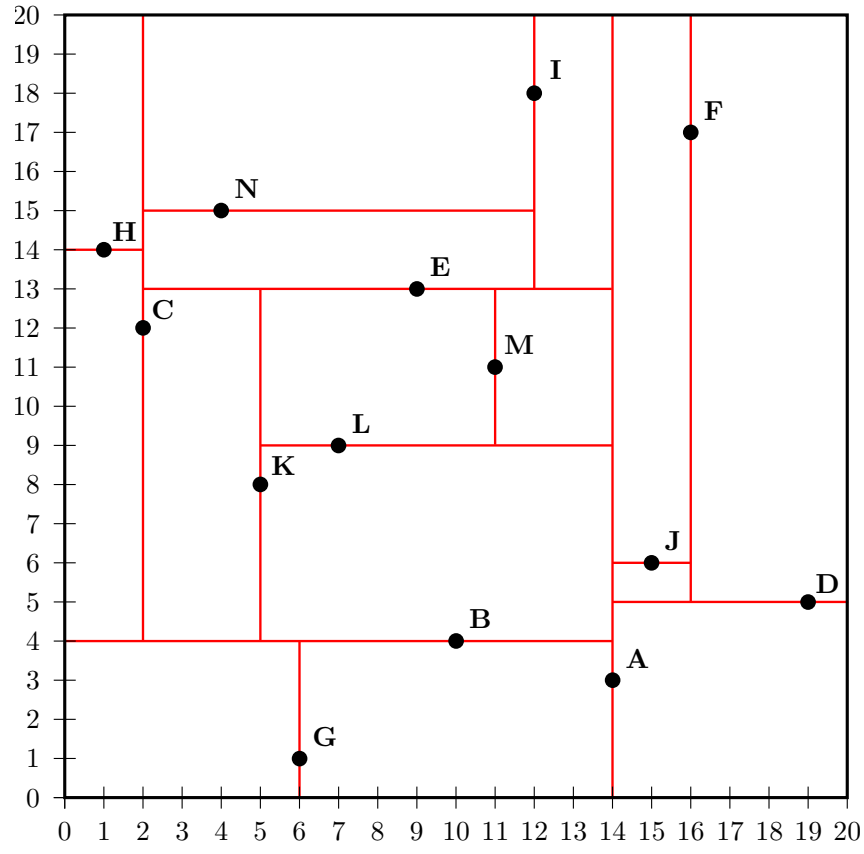
Example: K-d tree and induced partition



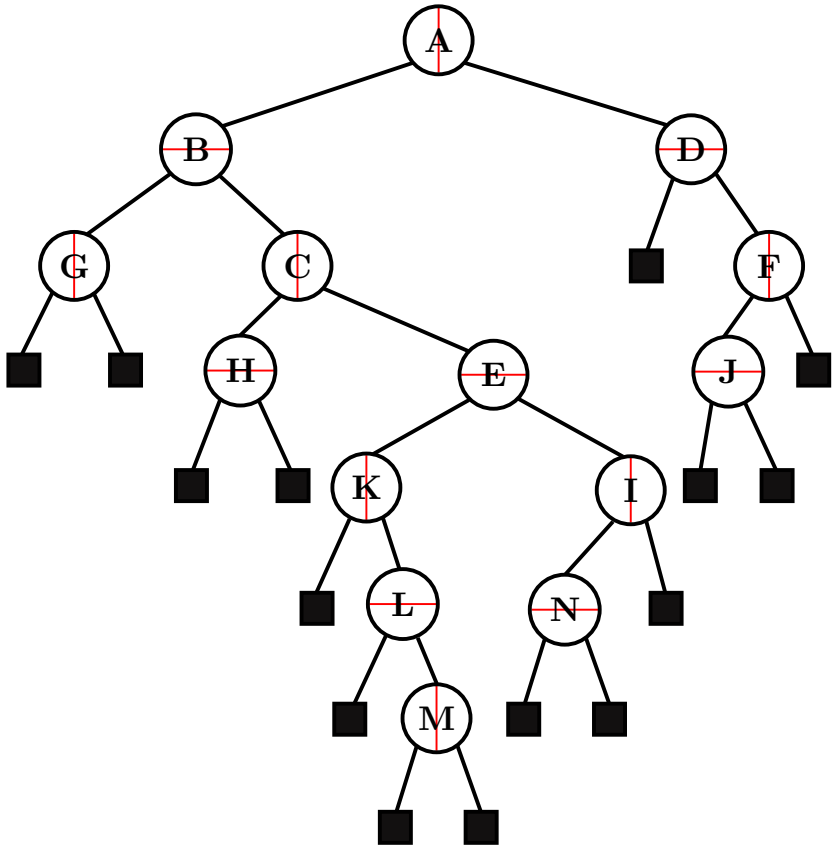


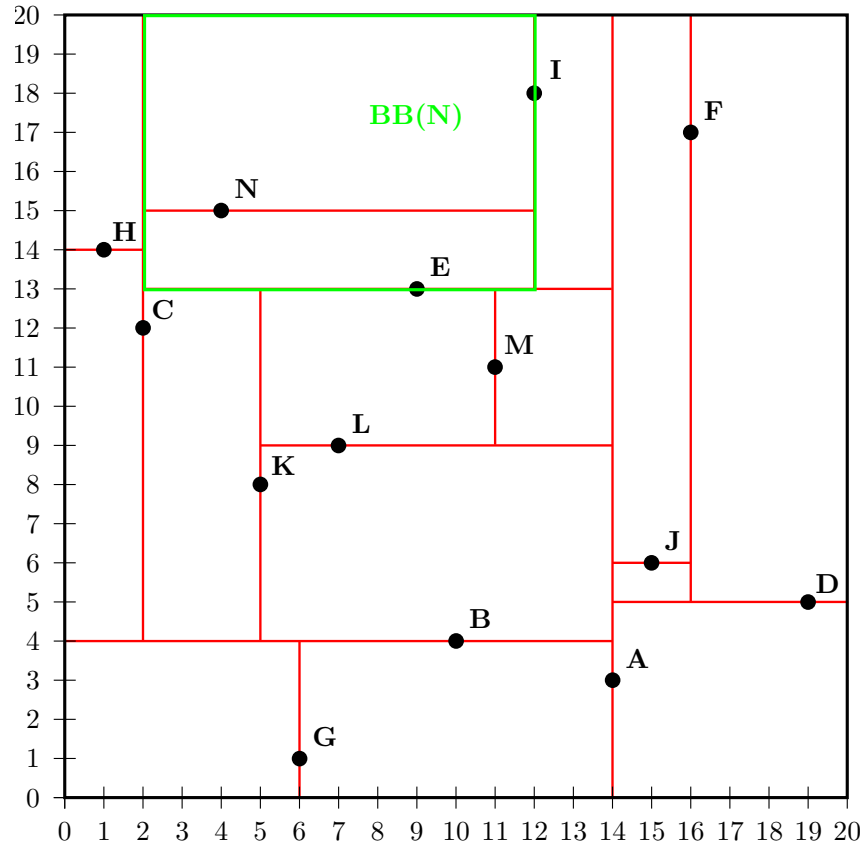
Example: K-d tree and induced partition





Example: K-d tree and induced partition

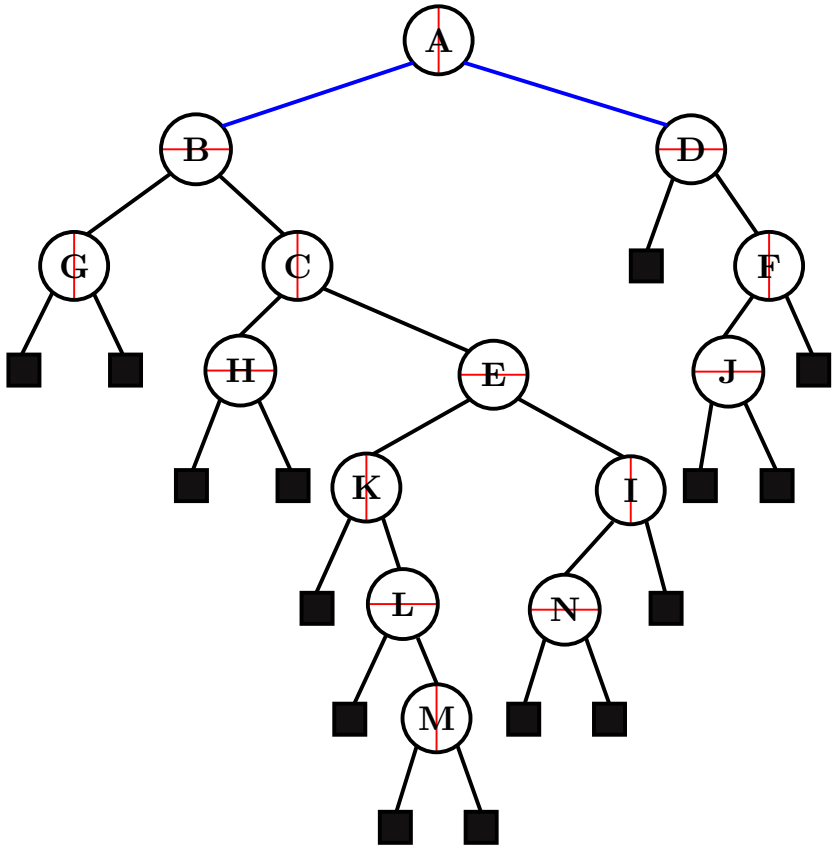


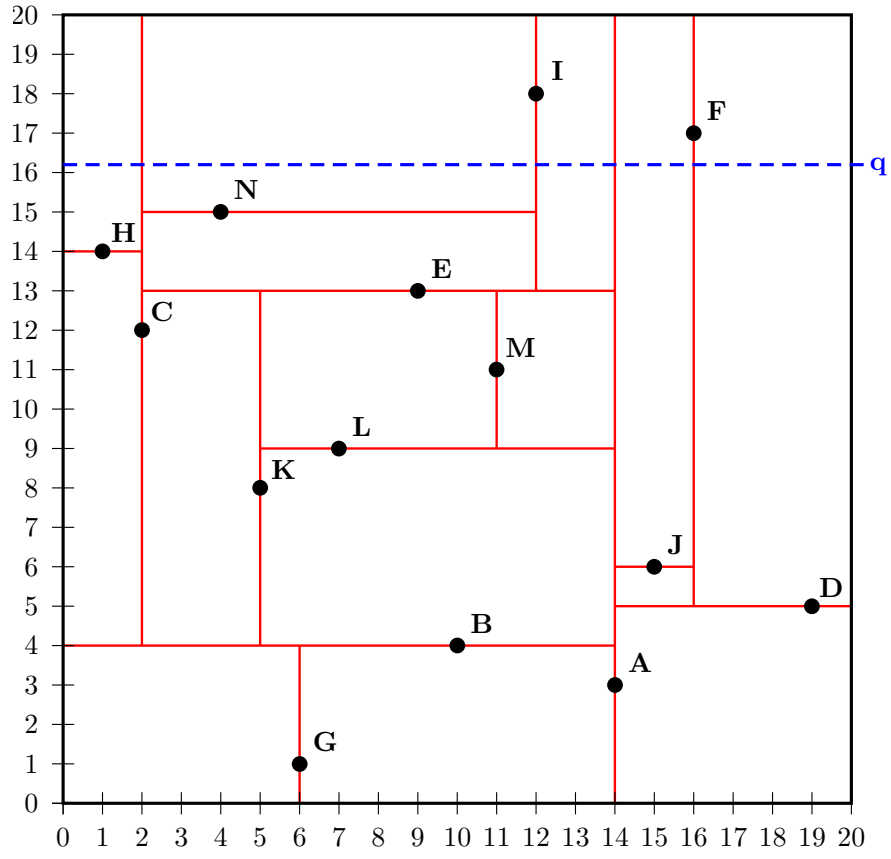


PM algorithm

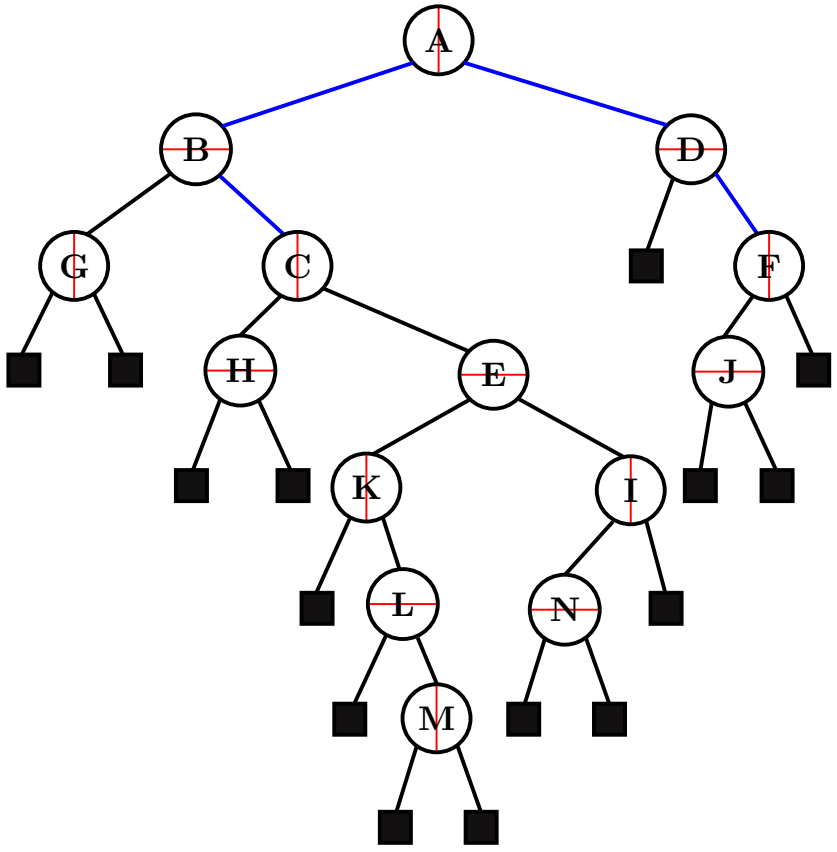
procedure PARTIALMATCH($\mathbf{q}$, T)
  if $T = \square$ **then return**
  $\mathbf{x} \leftarrow$ key of the root of T
  if $\mathbf{x}$ matches $\mathbf{q}$ **then**
    Report $\mathbf{x}$
  $i \leftarrow$ discriminant coordinate of the root of T
  if $q_i = *$ **then**
    PARTIALMATCH($\mathbf{q}$, left subtree of T)
    PARTIALMATCH($\mathbf{q}$, right subtree of T)
  else
    if $q_i \leq x_i$ **then**
      PARTIALMATCH($\mathbf{q}$, left subtree of T)
    else
      PARTIALMATCH($\mathbf{q}$, right subtree of T)

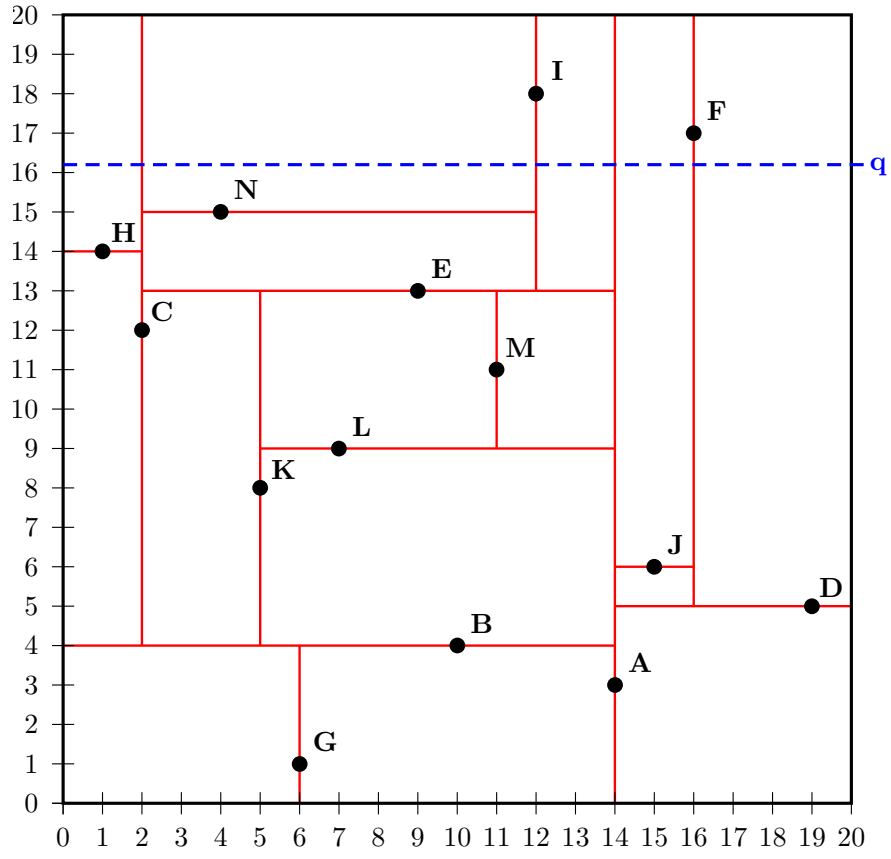
Example: PM query in a K-d tree



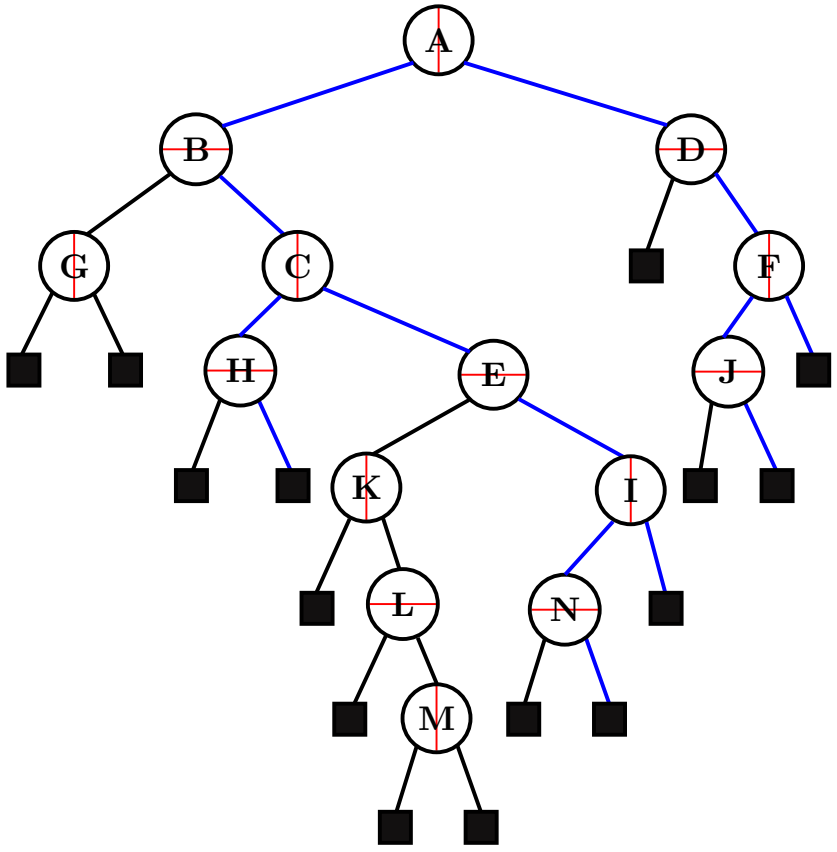


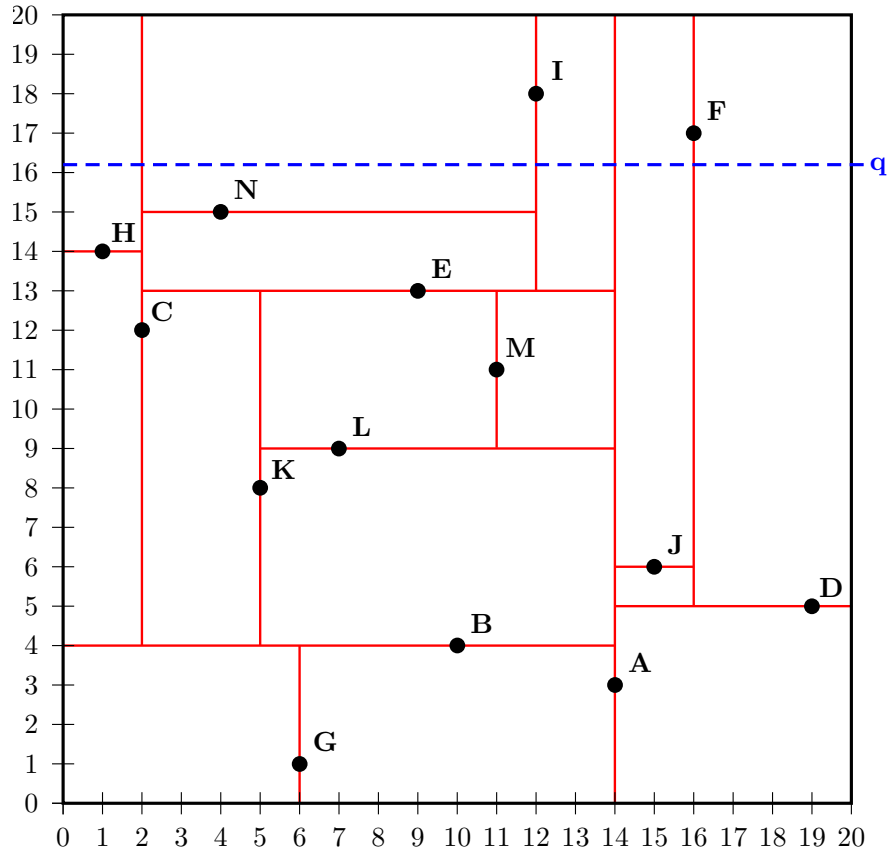
Example: PM query in a K-d tree





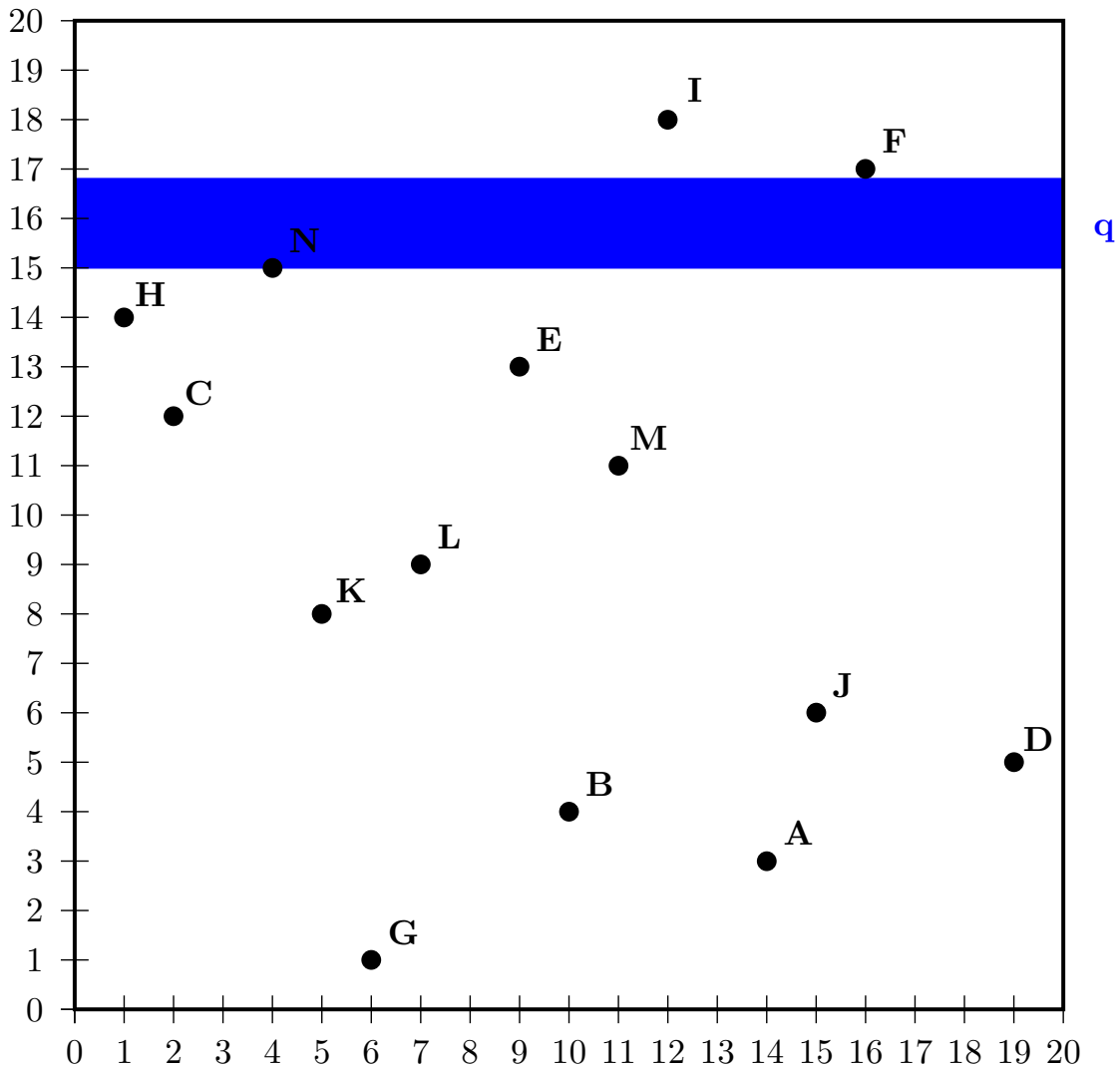
Example: PM query in a K-d tree





Cost of the PM algorithm

- The cost of the PM algorithm is measured as the number of visited nodes in the corresponding tree.
- If $\mathbf{r(q)} = \mathbf{r(q')}$ then cost of q = cost of q'.



Cost of the PM algorithm

The actual values x_i of the coordinates of the points in F are not relevant for the cost, only their relative ranks.

The cost depends on:

- The kind of hierarchical data structure under consideration.
- The probabilistic model from which data points are generated.

Probabilistic model: random data structures

Assumption:

The hierarchical multidimensional DS are built with the same probability from any of the $n!^K$ possible input sequences.

The cost of PM queries

In previous literature two different random variables have been analyzed:

- The cost of Fixed PM queries ($\mathcal{P}_{n,\mathbf{r}}$)
- The cost of the Randomized PM algorithm ($\mathcal{P}'_{n,\mathbf{u}}$)

We can also consider two different random variables:

- The cost of Random PM queries ($\hat{\mathcal{P}}_{n,\mathbf{u}}$)
- The cumulative cost of Fixed PM queries ($\tilde{\mathcal{P}}_{n,\mathbf{u}}$)

$$\mathbb{E}\{\mathcal{P}'_{n,\mathbf{u}}\} = \mathbb{E}\{\hat{\mathcal{P}}_{n,\mathbf{u}}\} = \mathbb{E}\{\tilde{\mathcal{P}}_{n,\mathbf{u}}\}/(n+1)^s$$

Random PM queries

Given a pattern $\mathbf{u}$, we define:

$$R_{n,\mathbf{u}} = \left\{ \mathbf{r} = (r_0, ..., r_{K-1}) \middle| r_i = * \text{ if } u_i = * \right.$$
$$\left. \text{and } 0 \leq r_i \leq n \text{ if } u_i = S \right\}$$

then the random variable that represents the cost of a random PM query is $\hat{\mathcal{P}}_{n,\mathbf{u}} := \mathcal{P}_{n,\mathcal{R}}$, where $\mathcal{R}$ is taken uniformly at random among the $(n+1)^s$ elements of $R_{n,\mathbf{u}}$.

Higher order moments of $\hat{\mathcal{P}}_{n,\mathbf{u}}$ can be easily obtained from the higher order moments of $\mathcal{P}_{n,\mathbf{r}}$.

Cumulative PM queries

The cumulative PM query is the sum of the fixed PM queries:

$$\tilde{\mathcal{P}}_{n,\mathbf{u}} = \sum_{\mathbf{r} \in R_{n,\mathbf{u}}} \mathcal{P}_{n,\mathbf{r}}$$

Higher order moments of $\tilde{\mathcal{P}}_{n,\mathbf{u}}$ would not be easy to obtain since one should know the co-variances of $\mathcal{P}_{n,\mathbf{r}}$ and $\mathcal{P}_{n,\mathbf{r}'}$.

Randomized PM algorithm

procedure RANDPARTIALMATCH($\mathbf{u}$, T)
  if $T = \square$ **then return**
  $\mathbf{x} \leftarrow$ key of the root of T
  $i \leftarrow$ discriminant coordinate of the root of T
  if $u_i = *$ **then**
    RANDPARTIALMATCH($\mathbf{u}$, left subtree of T)
    RANDPARTIALMATCH($\mathbf{u}$, right subtree of T)
  else
    { BB($\mathbf{x}$)$= [a_0, b_0] \times \cdots \times [a_{K-1}, b_{K-1}]$ }
    Generate a value $q_i \backsim$ Uniform (a_i, b_i).
    if $q_i \leq x_i$ **then**
      RANDPARTIALMATCH($\mathbf{u}$, left subtree of T)
    else
      RANDPARTIALMATCH($\mathbf{u}$, right subtree of T)

Randomized PM queries

$\mathcal{P}'_{n,\mathbf{u}} = \mathcal{P}'_n$, the cost of the randomized PM query, follows the distributional equation:

$$\mathcal{P}'_n = 1+\mathbf{1}_{\mathcal{S}}(\mathbf{1}_{V\leq\mathcal{U}}\mathcal{P}^{(1)}_{\mathcal{U}}+\mathbf{1}_{\mathcal{V}>\mathcal{U}}\mathcal{P}^{(2)}_{n-1-\mathcal{U}})+(1-\mathbf{1}_{\mathcal{S}})(\mathcal{P}^{(3)}_{\mathcal{U}}+\mathcal{P}^{(4)}_{n-1-\mathcal{U}})(*)$$

where:

- $\mathcal{S}$ follows a Bernoulli distribution with $p = s/K$ (indicates if the root discriminates by a specified coordinate), and
- $\mathcal{U}$ is Discrete Uniform $(0, n-1)$ (standing for the size of the left subtree),
- $\mathcal{V}$ is Discrete Uniform $(0, n)$ (marking whether the PM algorithm follows the left or the right subtrees),
- $\mathcal{P}^{(1)}, \mathcal{P}^{(2)}, \mathcal{P}^{(3)}, \mathcal{P}^{(4)}$ are independent copies of $\mathcal{P}'$.
- $\mathcal{S}, \mathcal{U}, \mathcal{V}$ and $\mathcal{P}', \mathcal{P}^{(1)}, \mathcal{P}^{(2)}, \mathcal{P}^{(3)}, \mathcal{P}^{(4)}$ are independent.

Some Previous Work

- Since the mid-80s there have been a substantial number of papers around the analysis of PM queries in many different multidimensional DS, e.g.,
  Flajolet and Puech (1986), Cunto, L. and Flajolet (1989), Flajolet, Gonnet, Puech and Robson (1993), Flajolet, Labelle, Laforest and Salvy (1995), Labelle and Laforest (1995), W. Schachinger (1995, 2000, 2004), Duch, Estivill-Castro and Martínez (1998), Devroye, Jabbour and Zamora-Cura (2000), Neininger (2000), Martinez, Panholzer and Prodinger (2001), Neininger and Rüschendorf (2001), Chern and Hwang (2003, 2006), ...
- Most of these papers study the expected cost of random PM queries (actually, the expected cost of RANDPARTIALMATCH), some consider the variance and distribution of the cost of RANDPARTIALMATCH.

Some Previous Work

- In the last few years, we have made considerable progress in the analysis of the cost of fixed PM queries $\mathcal{P}_{n,\mathbf{r}}$. From there it is inmediate to derive results for $\hat{\mathcal{P}}_{n,\mathbf{u}}$, the cost of a random PM query with pattern $\mathbf{u}$:
  Curien and Joseph (2011), Duch, Jiménez and Martínez (2012), Broutin, Neininger and Sulzbach (2013), Duch, Martínez and L. (2016), ...
- In particular, for $n \to \infty$, and $\mathbf{x} = \lim_{n\to\infty} \frac{1}{n}\mathbf{r}$, such that $0 < x_i < 1$ for all $x_i \neq *$, we have

$$\frac{\mathcal{P}_{n,\mathbf{r}}}{\beta n^{\alpha}} \xrightarrow{\mathcal{D}} h(\mathbf{x})\hat{\mathcal{P}}, \qquad (*)$$

with

$$h(\mathbf{x}) = \kappa \cdot \left(\prod_{i:x_i\neq *} x_i(1-x_i) \right)^{\alpha/2} \quad \text{and} \quad \frac{\hat{\mathcal{P}}_{n,\mathbf{u}}}{\beta n^{\alpha}} \xrightarrow{\mathcal{D}} \hat{\mathcal{P}}.$$

Some Previous Work

- This has been shown for 2-d quadtrees, standard K-d trees, and relaxed K-d trees.
- In (Duch, Martinez, L., 2016) we conjectured that this was true for other DS when $\alpha > 1 - s/K$.
- A similar result should hold for relaxed K-dt trees (a locally balanced variant of relaxed K-d trees), but we have proved that in this case

$$h(\mathbf{x}) \neq \kappa \cdot \left(\prod_{i:x_i \neq *} x_i(1-x_i) \right)^{\alpha/2}$$

(Duch and L., 2017).

Some Open Problems & Future Directions

- We are still interested in deriving "specific" results (distribution, expected value, other moments, ...) for particular DS, e.g., for the expected cost of a random PM query in relaxed K-dt trees we have the exponent $\alpha = \alpha(s, K, t)$, but no closed formula is yet known for $\beta = \beta(s, K, t)$.
- Another example is finding $h(\mathbf{x})$ for standard and relaxed K-dt trees , we have the linear ODE satisfied by h, for any t, but we've been unable to explicitly solve it, even for special cases, e.g., $t = 1$.
- We would like to have suitable "combinatorial" descriptions of several multidimensional DS, e.g., squarish K-d trees, median K-d trees, and then find closed formulas for β (random PM) or $h(\mathbf{x})$ (fixed PM).

Some Open Problems & Future Directions

- However our focus is in finding general results that apply in a wide variety of multidimensional DS, rather than finding specific results.
- We would like to provide general conditions under which

$$\frac{\mathcal{P}_{n,\mathbf{r}}}{\beta n^{\alpha}} \xrightarrow{\mathcal{D}} h(\mathbf{x})\hat{\mathcal{P}}, \qquad n \to \infty,$$

holds; on the other hand, preliminary results and experiments suggest that if $\alpha = 1 - s/K$ (e.g., squarish K-d trees, K-d tries, ...) then we have

$$\frac{\mathcal{P}_{n,\mathbf{r}}}{\beta n^{\alpha}} \xrightarrow{\mathcal{D}} \hat{\mathcal{P}}, \qquad n \to \infty,$$

with $\hat{\mathcal{P}}$ the limit distribution of the cost of random PM. Can we prove this without going on a case-by-case basis??

Some Open Problems & Future Directions

- Quad-Kd trees generalize several families of multidimensional DS, for instance all variants of K-d trees and quadtrees are particular instances.
- We have results on the expected cost of random PM for a large subclass of quad-Kd trees (Duch, L. and Martinez, 2016) which allow us to study, for instance, how α changes as we turn the "knob" from $p = 0$ (relaxed K-d trees) to $p = 1$ (quadtrees).
- We would be interested extending the subclass of quad-Kd trees on which our results apply and derive more results about the cost of PM queries on these trees, going beyond the expected cost of random PM queries.

This is the end, thank you very much for your attention!