An Optimal Ancestry Labeling Scheme

Pierre Fraigniaud Amos Korman[1]

CNRS and University Paris Diderot

[1] Speaker

Outline

Informative Labeling Scheme

Why should we fight for constants?

Optimal ancestry-labeling scheme

Small universal posets

Conclusion

Informative Labeling scheme

Graph representations:

- traditional: names given to the nodes serve merely as pointers to entries in a data structure
- informative labeling: mechanism for assigning short, yet informative, names to nodes (Kannan, Naor, Rudich [STOC '88])

General objective

To assign labels to nodes in such a way that allows one to infer information regarding any two nodes *directly from their labels*.

Main quality measure

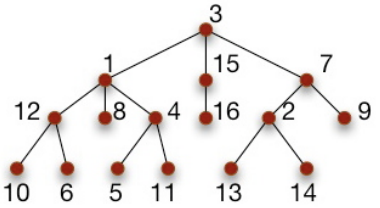
Label size = number of bits used to form the labels

Example 1: adjacency in trees

Input: tree T

Example 1: adjacency in trees

Input: tree T



1. Give distinct IDs to the nodes, between 1 and n
2. Root T at an arbitrary vertex

$$L(u) = (\mathrm{ID}(u), \mathrm{ID}(\mathrm{parent}(u)))$$

$$u \text{ and } v \text{ are adjacent} \iff u = \mathrm{parent}(v) \text{ or } v = \mathrm{parent}(u)$$

$$\text{Label size} = 2\lceil \log_2 n \rceil \text{ bits}$$

Informative Labeling Scheme

Let $\mathcal{P}$ be a boolean predicate defined on pairs of vertices for graphs in $\mathcal{F}$

Encoder (or marker) $\mathcal{M}$

Given $G \in \mathcal{F}$, $\mathcal{M}(G) = L$ where $L : V(G) \to \{0, 1\}^*$

Decoder $\mathcal{D}$

$\mathcal{D} : \{0, 1\}^* \times \{0, 1\}^* \to \{\text{true, false}\}$

For any $G \in \mathcal{F}$, and any $(u, v) \in V(G) \times V(G)$,

$$\mathcal{P}(u, v) = \text{true} \iff \mathcal{D}(L(u), L(v)) = \text{true}$$

Can be generalized to various types of functions (distance, connectivity, etc.), or tasks (e.g., routing).

Outline

Adjacency in trees

Definition

A graph $\mathcal{U}$ is universal for a graph family $\mathcal{F}$ if any $G \in \mathcal{F}$ is isomorphic to an induced subgraph of $\mathcal{U}$.

Theorem (Kannan, Naor, Rudich [STOC '88]**)**

There exists an adjacency labeling scheme for $\mathcal{F}$ with labels of at most k bits if and only if there exists a universal graph for $\mathcal{F}$ of order at most 2^k.

Adjacency in trees

Definition

A graph $\mathcal{U}$ is universal for a graph family $\mathcal{F}$ if any $G \in \mathcal{F}$ is isomorphic to an induced subgraph of $\mathcal{U}$.

Theorem (Kannan, Naor, Rudich [STOC '88])

There exists an adjacency labeling scheme for $\mathcal{F}$ with labels of at most k bits if and only if there exists a universal graph for $\mathcal{F}$ of order at most 2^k.

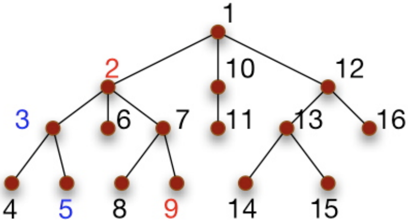
Adjacency: State of the art
$2\log n$ (Kannan, Naor, and Rudich [STOC '88])
$\log n + O(\log^* n)$ (Alstrup and Rauhe [FOCS '02])
$\Rightarrow$ universal graph of order $n2^{\log^* n}$

Example 2: ancestry in trees

Input: rooted tree

Example 2: ancestry in trees

Input: rooted tree



Give distinct DFS numbers to the nodes, between 1 and n

$$L(u) = (\text{DFS}(u), \text{DFS}(u_{max}))$$

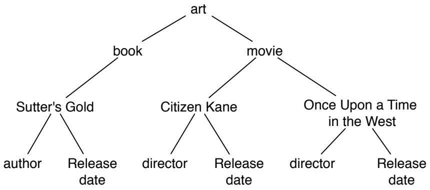
where u_{max} is the node with largest DFS number in the subtree rooted at u.

$$u \text{ is an ancestor of } v \iff \text{DFS}(v) \in [\text{DFS}(u), \text{DFS}(u_{max})]$$

$$\text{Label size} = 2\lceil \log_2 n \rceil \text{ bits}$$

XML trees

```
< art >
    < book >
        < Sutter's Gold >
            < author > Blaise Cendrars < /author >
            < Release > 1925 < /Release >
        < /Sutter's Gold >
    < /book >
    < movie >
        < Citizen Kane >
            < direct > Orson Wells < /direct >
            < Release > 1941 < /Release >
        < /Citizen Kane >
        < Once Upon a Time in the West >
            < direct > Sergio Leone < /direct >
            < Release > 1968 < /Release >
        < /Once Upon a Time in the West >
    < /movie >
< /art >
```



- Answer queries using the index labels only, without accessing the actual documents.
- A small improvement in the label size $\Rightarrow$ significant improvement in the performances of XML search engines.

State of the art: ancestry in trees

Ancestry
$2\log n$ (Kannan, Naor, and Rudich [STOC '88])
$\frac{3}{2}\log n + O(\log\log n)$ (Abiteboul, Kaplan, and Milo [SODA '01])
$\log n + O(\log n / \log\log n)$ (Thorup and Zwick [SPAA '01])
$\log n + O(\sqrt{\log n})$ (Alstrup and Rauhe [SODA '02])
$\log n + \Omega(\log\log n)$ (Alstrup, Bille and Rauhe [SODA '03])
$\log n + 2\log(\mathit{depth}) + O(1)$ (Fraigniaud and Korman, [SODA '10])
$\log n + O(\log\log n)$ (Fraigniaud and Korman, [STOC '10])

Outline

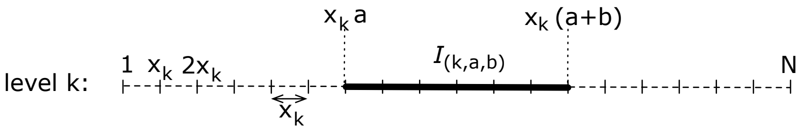
Interval containment

v ancestor of $u \iff I(u) \subseteq I(v)$

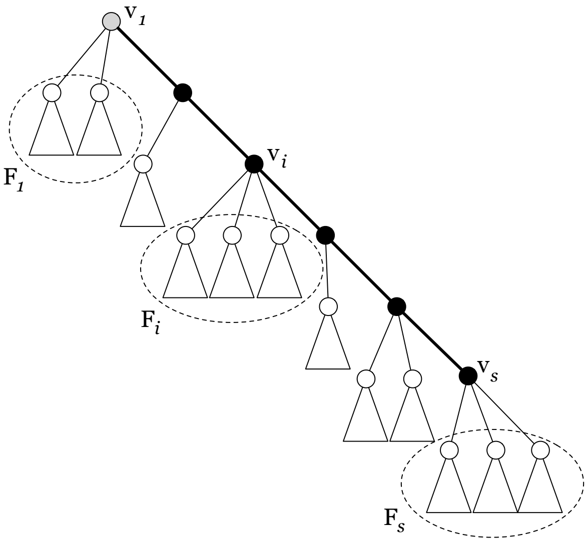
$2\log n$-scheme by Kannan, Naor, and Rudich use n^2 intervals.

We aim at using $n \log^c n$ intervals

We use intervals of the following form, for $k = 1, \ldots, \log n$:



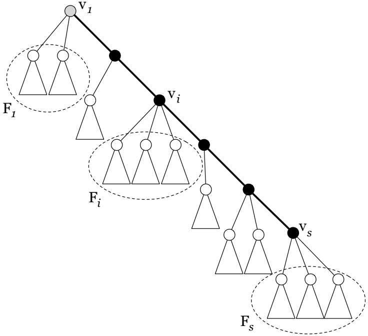
Spine decomposition



Nodes classified as either *heavy* or *apex*.

Trees with bounded spine decomposition depth $d = 0(1)$

Trees with bounded spine decomposition depth $d = 0(1)$



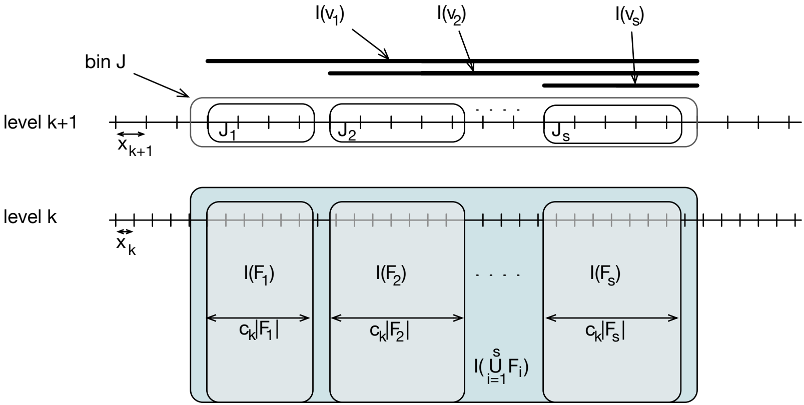
$\mathcal{F}(n, d)$ = forests with $\leq n$ nodes, and spine-decomposition depth $\leq d$.

We aim at using nd^2 intervals for $F \in \mathcal{F}(n, d)$

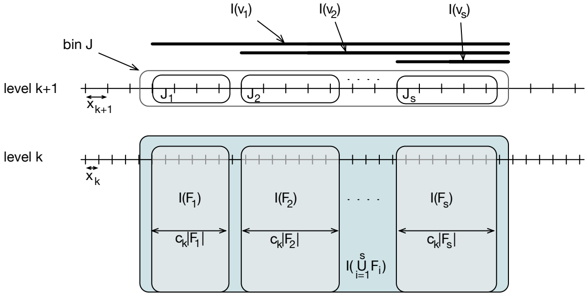
Induction of $k = \log n$

Difficult case: F containing a tree T of size larger than 2^k, i.e., $2^k < |T| \leq 2^{k+1}$.

General idea



Tuning of the parameters (1/3)



For $1 \leq i < s$, the length of $I(v_i)$ must satisfy

$$|I(v_i)| \approx c_k|F_i| + x_{k+1} + |I(v_{i+1})| \approx c_k(\sum_{j=i}^{s}|F_i|) + i \cdot x_{k+1}.$$

Bin J to be of length $|J| \approx c_k \cdot 2^{k+1} + (s+1) \cdot x_{k+1}$ suffices.

Tuning of the parameters (2/3)

Since $s \leq d$, we must have $|J|$ be approximately

$$c_{k+1} 2^{k+1} \;\approx\; c_k 2^{k+1} + d \cdot x_{k+1}$$

Choose the values of c_k so that:

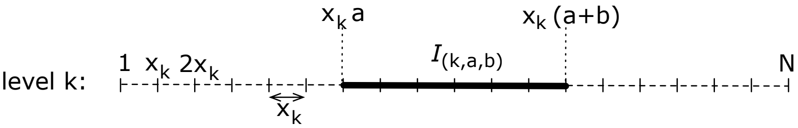
$$c_{k+1} - c_k \;\approx\; \frac{d \cdot x_{k+1}}{2^{k+1}}$$

We set

$$c_k \approx \sum_{j=1}^{k} \frac{1}{j^{1+\epsilon}}, \quad \text{and thus} \quad x_k \approx \frac{2^k}{d \cdot k^{1+\epsilon}}$$

Tuning of the parameters (3/3)

Let $A_k \approx N/x_k$ and $B_k \approx c_k 2^k/x_k$.



where $1 \leq a \leq A_k$ and $1 \leq b \leq B_k$.

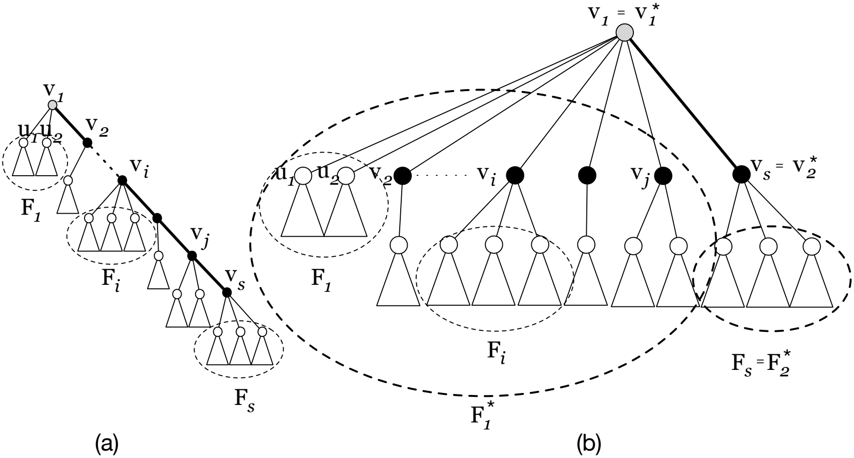
Thus, $N \approx c_{\log n} \cdot n = O(n)$.

The number of level-k intervals is

$$O(A_k \cdot B_k) = O(nd^2 k^{2(1+\epsilon)}/2^k),$$

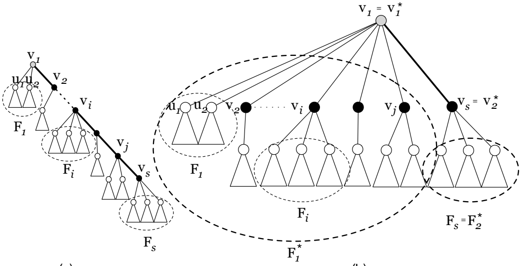
yielding a total of $O(nd^2)$ intervals, as desired.

The general case: uses the folding-decomposition



Ancestry preservation

DFS traversal in T that visits apex children first.
For any node u, let $\texttt{DFS}(u)$ be the DFS number of u.



(a) (b)

Lemma

Node v is an ancestor of u in T if and only if at least one of the following two conditions hold

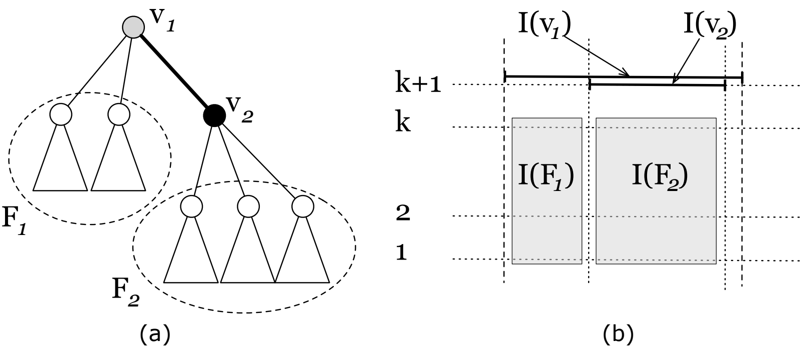
- **C1:** *v is an ancestor of u in T^*;*
- **C2:** $\texttt{APEX}(v)$ *is ancestor of u in T^* and* $\texttt{DFS}(v) < \texttt{DFS}(u)$.

Ordering the intervals

Lemma

Node v is an ancestor of u in T if and only if at least one of the following two conditions hold

- **C1:** *v is an ancestor of u in T^*;*
- **C2':** $\text{APEX}(v)$ *is ancestor of u in T^* and $I(v) \prec I(u)$.*

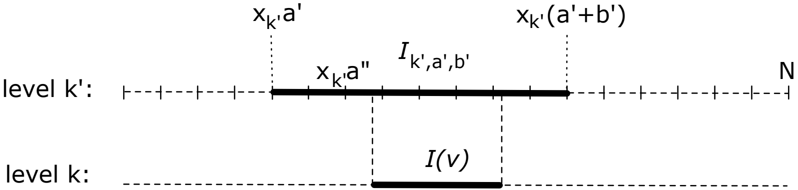


$$label(u) = (I(u), I(\text{APEX}(u)))$$

Compact encoding of $I(\text{APEX}(v))$

It is sufficient to encode:

- its level k'
- two shifts b'_{left} and b'_{right} in $[1, B_{k'}]$



Outline

Graph arboricity

The arboricity of a graph is the minimum number of forests into which its edges can be partitioned.

Corollary (Kannan, Naor, Rudich [STOC '88])

There exists an adjacency labeling scheme for the family of graphs with arboricity at most k with labels of at most $(k+1)\log n$ bits.

High level correspondence between:

adjacency/arboricity for graphs
and
ancestry/tree-dimension for posets

Partially ordered sets

Poset $(X, \leq)$

- reflexivity: $x \leq x$
- antisymmetry: ($x \leq y$ and $y \leq x$) $\quad \Rightarrow \quad x = y$
- transitivity: ($x \leq y$ and $y \leq z$) $\quad \Rightarrow \quad x \leq z$

$(X, \leq')$ is an extension of $(X, \leq)$ if:

$$\forall x, y \in X, \;\; x \leq y \Rightarrow x \leq' y$$

The dimension of a poset $(X, \leq)$ is the smallest number of linear (i.e., total order) extensions of $(X, \leq)$ the intersection of which gives rise to $(X, \leq)$.

Universal posets

A poset $(X, \leq_X)$ contains a poset $(Y, \leq_Y)$ as an induced suborder if there exists an injective mapping $\phi : Y \to X$ such that for any two elements $a, b \in Y$:

$$a \leq_Y b \iff \phi(a) \leq_X \phi(b).$$

Definition

A poset $(\mathcal{U}, \leq)$ is called universal for a family of posets $\mathcal{F}$ if $(\mathcal{U}, \leq)$ contains every poset in $\mathcal{F}$ as an induced suborder.

The size of a universal posets

Remark

The smallest size of a universal poset for the family of n-element posets with dimension at most k is at most n^k.

Theorem (Alon and Scheinerman [Order 1988])

The number of n-element posets of dimension k is at least $n^{n(k-o(1))}$.

Corollary

A universal poset for the family of all n-element posets with dimension at most k has number of elements at least $n^{k-o(1)}$.

Tree dimension

Definition

A poset $(X, \leq)$ is a tree if, for every pair x and y of incomparable elements in X, there does not exist an element $z \in X$ such that $x \leq z$ and $y \leq z$.

The tree-dimension of a poset $(X, \leq)$ is the smallest number of tree extensions of $(X, \leq)$ the intersection of which gives rise to $(X, \leq)$.

Universal posets for tree-dimension k

$$\text{tree-dim} \leq \text{dim} \leq 2 \cdot \text{tree-dim}$$

Thus, the smallest size of a universal poset for the family of all n-element posets with tree-dimension at most k is:

- at least $n^{k-o(1)}$, and
- at most n^{2k}.

Theorem (Fraigniaud and Korman [STOC 2010])

For every integer k, there exists a universal poset of size $O(n^k \log^{4k} n)$ for the family of the n-element posets of tree-dimension k.

Outline

Further work

Open problem

- Is the size of a smallest universal graph for trees with at most n nodes linear in n?
- Recall that we know it is of size at most $n2^{O(\log^* n)}$.

Randomization

- Randomized ancestry labeling schemes (1-sided error).
- Tradeoffs can be established for adjacency [Fraigniaud and Korman, SPAA 2009].

Generalization to "dynamic network"

- What is a dynamic graph?
- What type of complexity measure?

Further work

Open problem

- Is the size of a smallest universal graph for trees with at most n nodes linear in n?
- Recall that we know it is of size at most $n2^{O(\log^* n)}$.

Randomization

- Randomized ancestry labeling schemes (1-sided error).
- Tradeoffs can be established for adjacency [Fraigniaud and Korman, SPAA 2009].

Generalization to "dynamic network"

- What is a dynamic graph?
- What type of complexity measure?

Thank You!