A Comparison of Empirical Tree Entropies

Danny Hucke, Markus Lohrey, and Louisa Seelbach Benkner
University of Siegen

SPIRE 2020

Empirical Entropy for Strings

- Well-established measure for the compressibility of a single string

Empirical Entropy for Strings

- Well-established measure for the compressibility of a single string
- Introduced by Kosaraju and Manzini in 1999

Empirical Entropy for Strings

- Well-established measure for the compressibility of a single string
- Introduced by Kosaraju and Manzini in 1999
- Used to analyze the performance of various string compression algorithms (SLPs, BWT,...)

Empirical Entropy for Strings

- Well-established measure for the compressibility of a single string
- Introduced by Kosaraju and Manzini in 1999
- Used to analyze the performance of various string compression algorithms (SLPs, BWT,...)
- Expected uncertainty about a symbol of the string, given the k preceding characters

Empirical Entropy for Strings

- Well-established measure for the compressibility of a single string
- Introduced by Kosaraju and Manzini in 1999
- Used to analyze the performance of various string compression algorithms (SLPs, BWT,...)
- Expected uncertainty about a symbol of the string, given the k preceding characters
- Let $s \in \Sigma^*$ and let $\#_x(s)$ denote the number of occurrences of character $x \in \Sigma$ in s
- Let $\alpha \in \Sigma^k$ for $k > 0$ and let s_α denote the string whose i^{th} character is the character in s following the i^{th} occurrence of α in s
- (Unnormalized) empirical entropy of order k of s:

$$H_k(s) = \sum_{\alpha \in \Sigma^k} \sum_{x \in \Sigma} \#_x(s_\alpha) \log_2 \left(\frac{|s_\alpha|}{\#_x(s_\alpha)} \right)$$

Empirical Entropy for Strings

- Let $s \in \Sigma^*$ and let $\#_x(s)$ denote the number of occurrences of character $x \in \Sigma$ in s
- Let $\alpha \in \Sigma^k$ for $k > 0$ and let s_α denote the string whose i^{th} character is the character in s following the i^{th} occurrence of α in s
- (Unnormalized) empirical entropy of order k of s:

$$H_k(s) = \sum_{\alpha \in \Sigma^k} \sum_{x \in \Sigma} \#_x(s_\alpha) \log_2 \left(\frac{|s_\alpha|}{\#_x(s_\alpha)} \right)$$

- Example: Second order empirical entropy of

s = bananabanana

Empirical Entropy for Strings

- Let $s \in \Sigma^*$ and let $\#_x(s)$ denote the number of occurrences of character $x \in \Sigma$ in s
- Let $\alpha \in \Sigma^k$ for $k > 0$ and let s_α denote the string whose i^{th} character is the character in s following the i^{th} occurrence of α in s
- (Unnormalized) empirical entropy of order k of s:

$$H_k(s) = \sum_{\alpha \in \Sigma^k} \sum_{x \in \Sigma} \#_x(s_\alpha) \log_2 \left(\frac{|s_\alpha|}{\#_x(s_\alpha)} \right)$$

- Example: Second order empirical entropy of

s = bananabanana

α_1=ba, s_{α_1} =nn;

Empirical Entropy for Strings

- Let $s \in \Sigma^*$ and let $\#_x(s)$ denote the number of occurrences of character $x \in \Sigma$ in s
- Let $\alpha \in \Sigma^k$ for $k > 0$ and let s_α denote the string whose i^{th} character is the character in s following the i^{th} occurrence of α in s
- (Unnormalized) empirical entropy of order k of s:

$$H_k(s) = \sum_{\alpha \in \Sigma^k} \sum_{x \in \Sigma} \#_x(s_\alpha) \log_2 \left(\frac{|s_\alpha|}{\#_x(s_\alpha)} \right)$$

- Example: Second order empirical entropy of

s = bananabanana

α_1=ba, s_{α_1} =nn; α_2 = an, s_{α_2} =aaaa

Empirical Entropy for Strings

- Let $s \in \Sigma^*$ and let $\#_x(s)$ denote the number of occurrences of character $x \in \Sigma$ in s
- Let $\alpha \in \Sigma^k$ for $k > 0$ and let s_α denote the string whose i^{th} character is the character in s following the i^{th} occurrence of α in s
- (Unnormalized) empirical entropy of order k of s:

$$H_k(s) = \sum_{\alpha \in \Sigma^k} \sum_{x \in \Sigma} \#_x(s_\alpha) \log_2 \left(\frac{|s_\alpha|}{\#_x(s_\alpha)} \right)$$

- Example: Second order empirical entropy of

s =bananabanana

α_1=ba, s_{α_1} =nn; α_2 = an, s_{α_2} =aaaa;
$\alpha_3 = na$, $s_{\alpha_3} = nbn$;

Empirical Entropy for Strings

- Let $s \in \Sigma^*$ and let $\#_x(s)$ denote the number of occurrences of character $x \in \Sigma$ in s
- Let $\alpha \in \Sigma^k$ for $k > 0$ and let s_α denote the string whose i^{th} character is the character in s following the i^{th} occurrence of α in s
- (Unnormalized) empirical entropy of order k of s:

$$H_k(s) = \sum_{\alpha \in \Sigma^k} \sum_{x \in \Sigma} \#_x(s_\alpha) \log_2 \left(\frac{|s_\alpha|}{\#_x(s_\alpha)} \right)$$

- Example: Second order empirical entropy of

s =bananabanana

α_1=ba, s_{α_1} =nn; α_2 = an, s_{α_2} =aaaa;
$\alpha_3 = na$, $s_{\alpha_3} = nbn$; $\alpha_4 = ab$, $s_{\alpha_4} = a$;

Empirical Entropy for Strings

- Let $s \in \Sigma^*$ and let $\#_x(s)$ denote the number of occurrences of character $x \in \Sigma$ in s
- Let $\alpha \in \Sigma^k$ for $k > 0$ and let s_α denote the string whose i^{th} character is the character in s following the i^{th} occurrence of α in s
- (Unnormalized) empirical entropy of order k of s:

$$H_k(s) = \sum_{\alpha \in \Sigma^k} \sum_{x \in \Sigma} \#_x(s_\alpha) \log_2 \left(\frac{|s_\alpha|}{\#_x(s_\alpha)} \right)$$

- Example: Second order empirical entropy of

s =bananabanana

α_1=ba, s_{α_1} =nn; α_2 = an, s_{α_2} =aaaa;
$\alpha_3 = na$, $s_{\alpha_3} = nbn$; $\alpha_4 = ab$, $s_{\alpha_4} = a$;

$$H_2(s) = 2 \log_2 \left(\frac{3}{2} \right) + 1 \log_2 \left(\frac{3}{1} \right)$$

Empirical Entropy for Trees

- Goal: Generalize concepts/results from information theory and data compression from strings to structured data like graphs/trees
- For strings: Basically one notion of empirical entropy
- For trees: Several notions of empirical tree entropy have been proposed in the last few years
- Trees: Compressibility depends on labels and structure, the degree/label of a node might depend on its label/degree/child-rank/ancestors' labels/siblings' labels/...

Empirical Entropy for Trees

There are plenty of notions of empirical entropy for trees:

Label Entropy H_k^ℓ
Ferragina et al., 2005

Empirical Entropy for Trees

There are plenty of notions of empirical entropy for trees:

Label Entropy H_k^ℓ
Ferragina et al., 2005

Degree Entropy $H^{\deg}$
Jansson et al., 2012

Empirical Entropy for Trees

There are plenty of notions of empirical entropy for trees:

Label Entropy H_k^{ℓ}
Ferragina et al., 2005

Degree Entropy $H^{\deg}$
Jansson et al., 2012

Degree-Label Entropy $H_k^{\deg,\ell}$
Ganczorz, 2020

Label-Degree Entropy $H_k^{\ell,\deg}$
Ganczorz, 2020

Empirical Entropy for Trees

There are plenty of notions of empirical entropy for trees:

Label Entropy H_k^{ℓ}
Ferragina et al., 2005

Degree Entropy $H^{\deg}$
Jansson et al., 2012

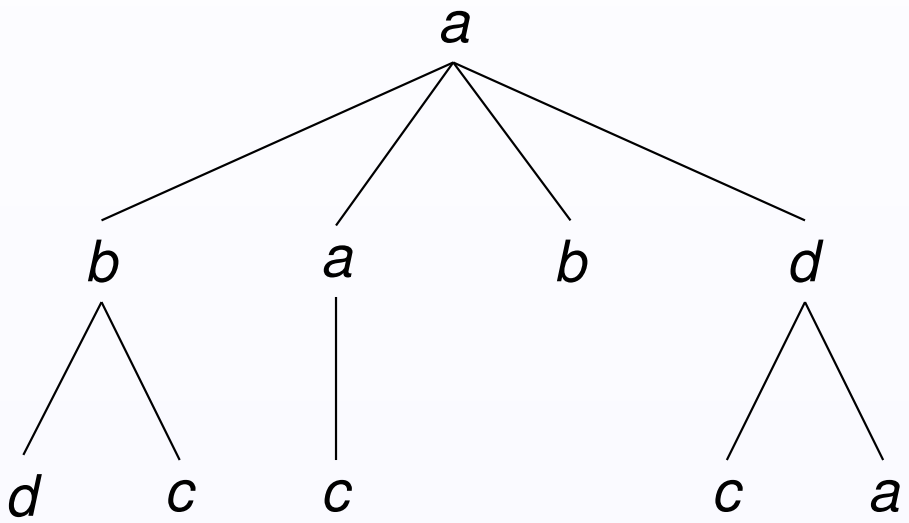
Degree-Label Entropy $H_k^{\deg,\ell}$
Ganczorz, 2020

Label-Degree Entropy $H_k^{\ell,\deg}$
Ganczorz, 2020

Label-Shape Entropy $H_k^{\ell s}$
Hucke et al., 2019

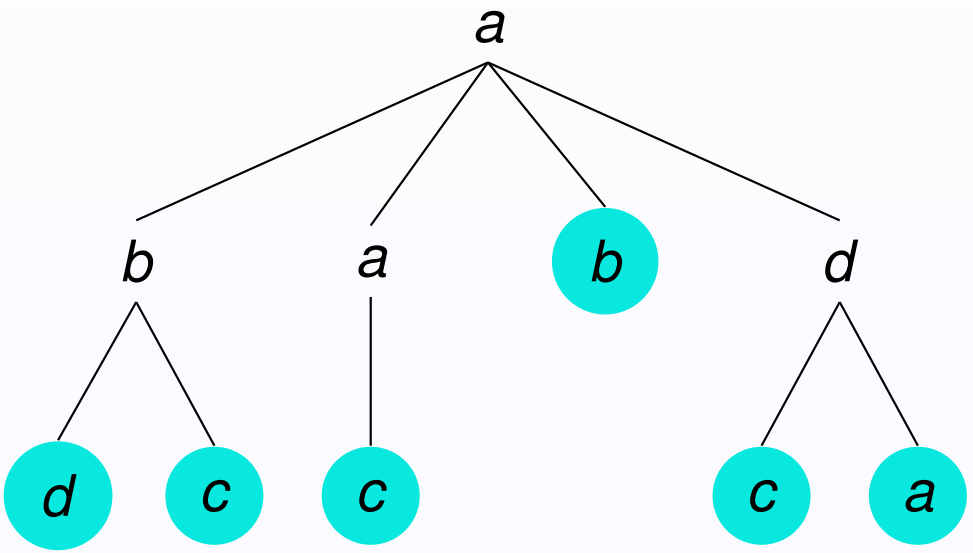
Degree Entropy $H^{\deg}$

- Defined by Jansson, Sadakane and Sung in 2012
- Zeroth order empirical entropy of the node degrees occurring in the tree



Degree Entropy $H^{\deg}$

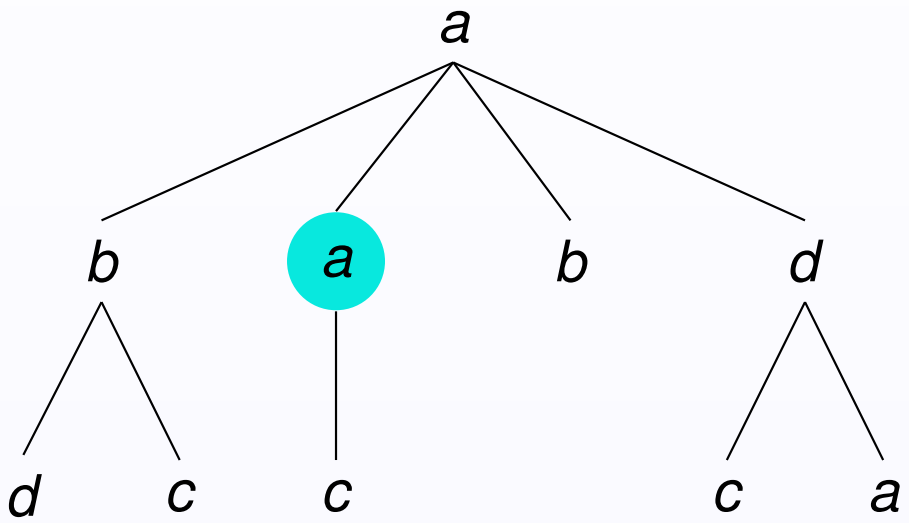
- Defined by Jansson, Sadakane and Sung in 2012
- Zeroth order empirical entropy of the node degrees occurring in the tree



- Degree 0: 6 nodes

Degree Entropy $H^{\deg}$

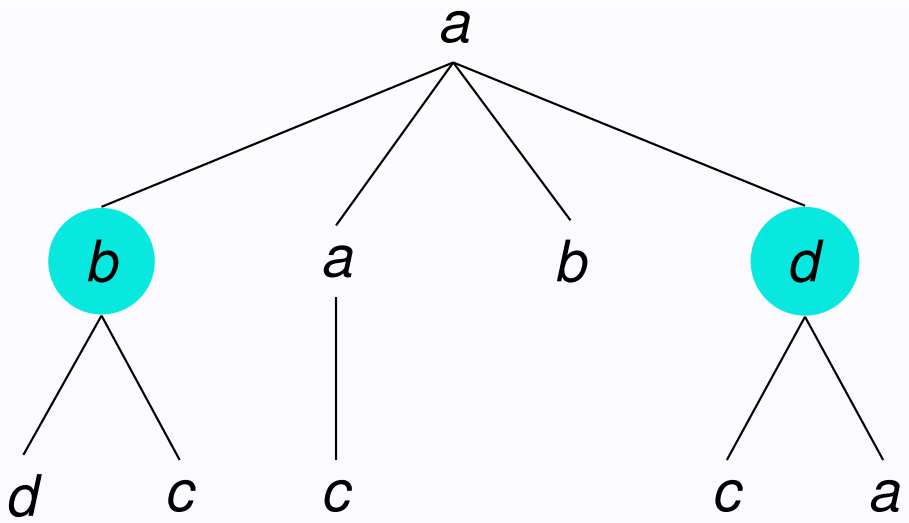
- Defined by Jansson, Sadakane and Sung in 2012
- Zeroth order empirical entropy of the node degrees occurring in the tree



- Degree 0: 6 nodes
- Degree 1: 1 node

Degree Entropy $H^{\deg}$

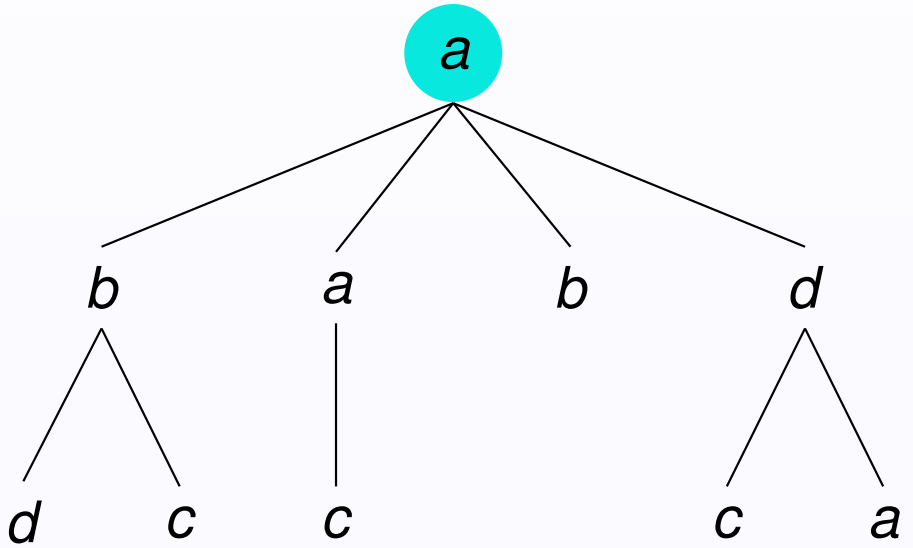
- Defined by Jansson, Sadakane and Sung in 2012
- Zeroth order empirical entropy of the node degrees occurring in the tree



- Degree 0: 6 nodes
- Degree 1: 1 node
- Degree 2: 2 nodes

Degree Entropy $H^{\deg}$

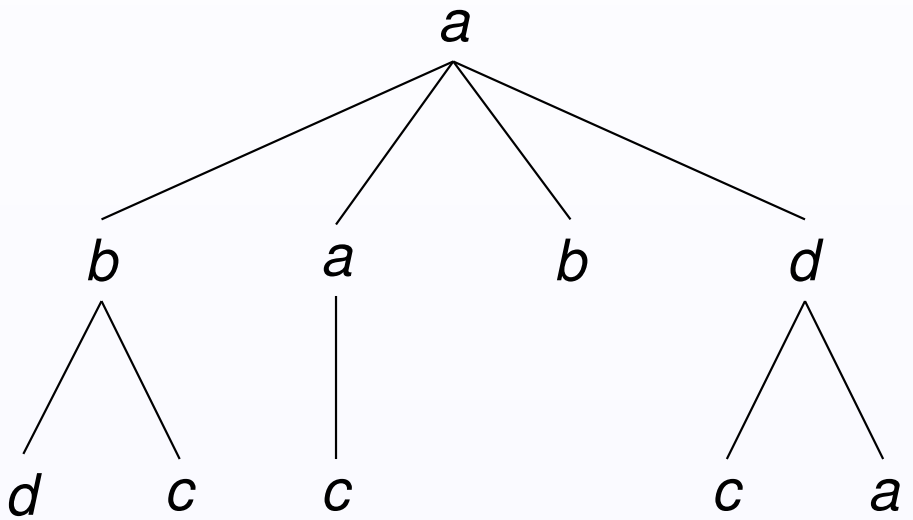
- Defined by Jansson, Sadakane and Sung in 2012
- Zeroth order empirical entropy of the node degrees occurring in the tree



- Degree 0: 6 nodes
- Degree 1: 1 node
- Degree 2: 2 nodes
- Degree 4: 1 node

Degree Entropy $H^{\deg}$

- Defined by Jansson, Sadakane and Sung in 2012
- Zeroth order empirical entropy of the node degrees occurring in the tree



- Degree 0: 6 nodes
- Degree 1: 1 node
- Degree 2: 2 nodes
- Degree 4: 1 node
- Total: 10 nodes

$$H^{\deg}(t) = 6 \log_2 \left(\frac{10}{6}\right) + 1 \log_2 \left(\frac{10}{1}\right) + 2 \log_2 \left(\frac{10}{2}\right) + 1 \log_2 \left(\frac{10}{1}\right)$$

$$\approx 15,7095...$$

Degree Entropy $H^{\deg}$

Theorem (Jansson, Sadakane, Sung, 2012.)

Let t be an unlabeled tree, then t can be represented in

$$H^{\deg}(t)$$

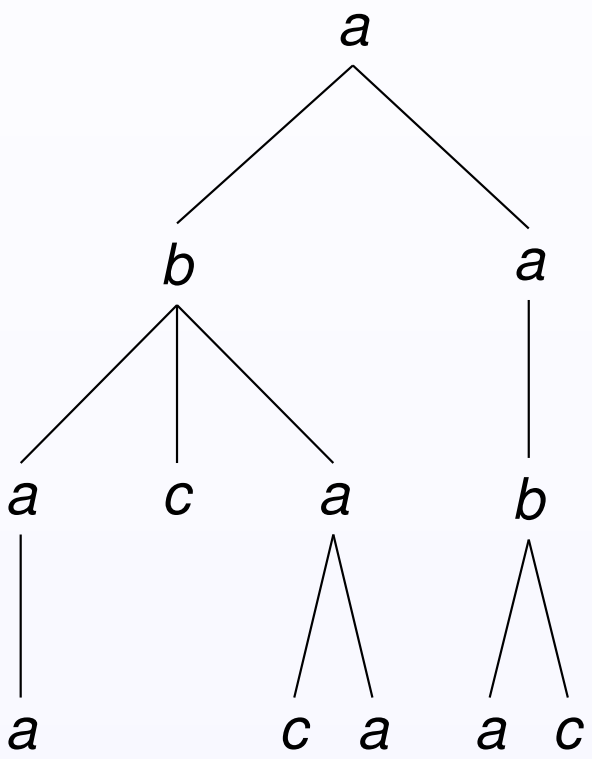
(plus lower-order term) many bits.

Label Entropy H_k^ℓ

- Introduced by Ferragina et al. in 2005
- Expected uncertainty about the label of a node v, given the k first labels on the path from v's parent to the root (its *k-label-history*)

Label Entropy H_k^ℓ

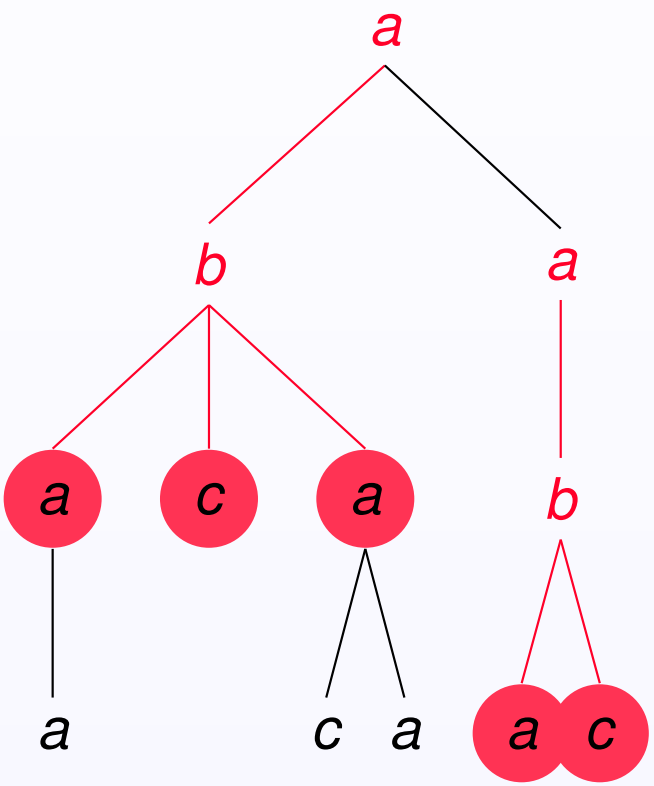
- Introduced by Ferragina et al. in 2005
- Expected uncertainty about the label of a node v, given the k first labels on the path from v's parent to the root (its *k-label-history*)



2nd-order label entropy:

Label Entropy H_k^ℓ

- Introduced by Ferragina et al. in 2005
- Expected uncertainty about the label of a node v, given the k first labels on the path from v's parent to the root (its *k-label-history*)

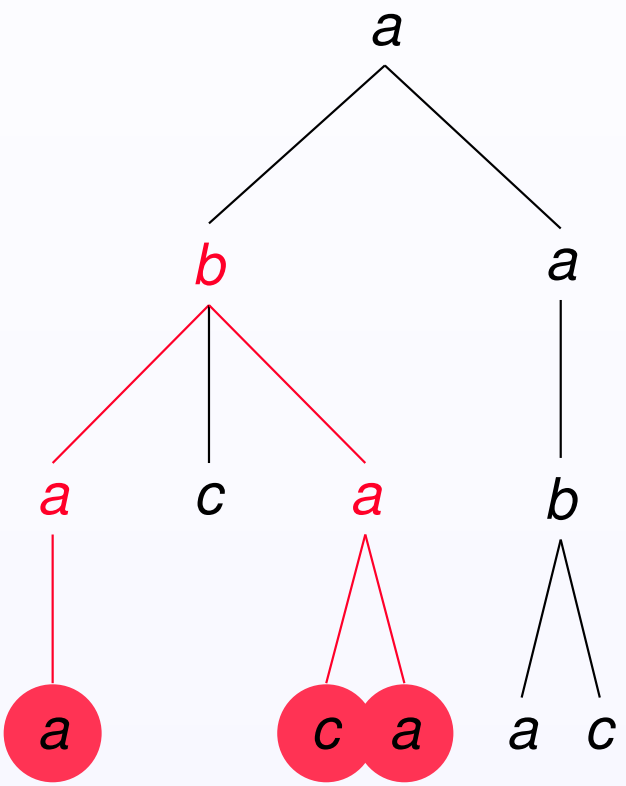


2nd-order label entropy:

- 2-history *ab*: 5 nodes total, 3 nodes labeled *a*, 2 nodes labeled *c*

Label Entropy H_k^ℓ

- Introduced by Ferragina et al. in 2005
- Expected uncertainty about the label of a node v, given the k first labels on the path from v's parent to the root (its *k-label-history*)

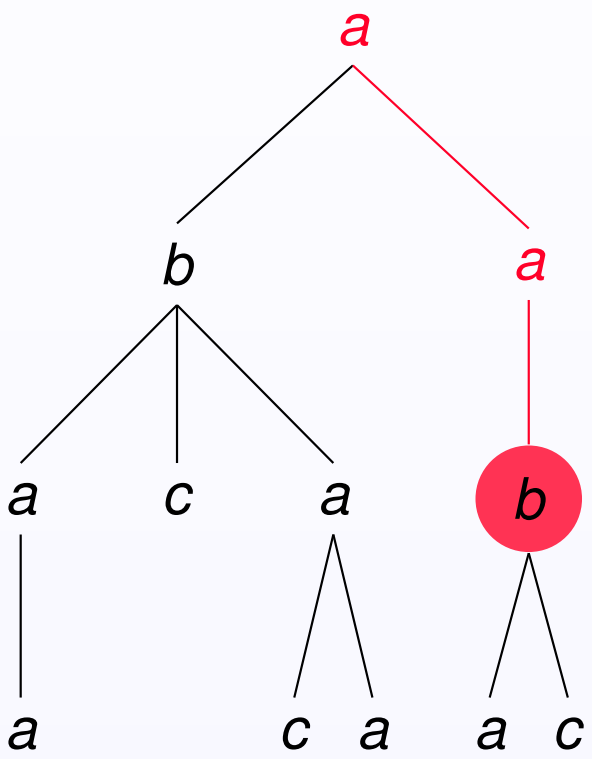


2nd-order label entropy:

- 2-history *ab*: 5 nodes total, 3 nodes labeled *a*, 2 nodes labeled *c*
- 2-history *ba*: 3 nodes total, 2 nodes labeled *a*, 1 node labeled *c*

Label Entropy H_k^ℓ

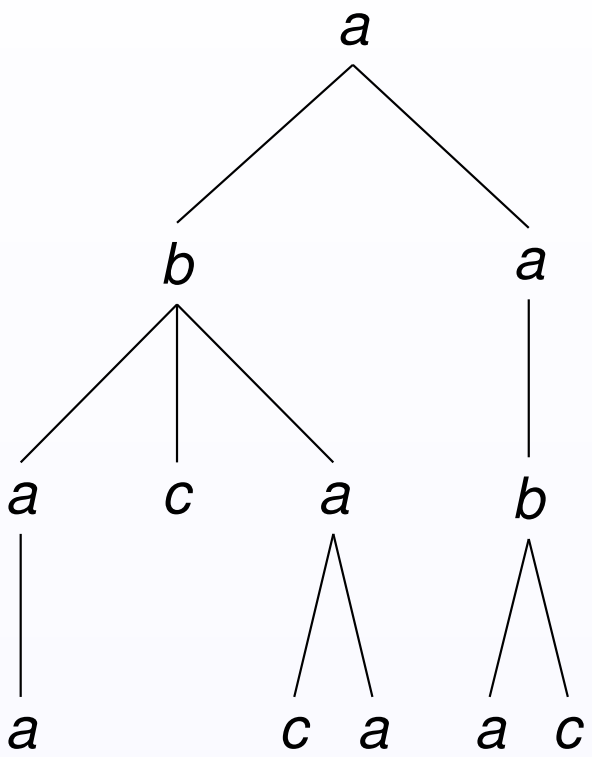
- Introduced by Ferragina et al. in 2005
- Expected uncertainty about the label of a node v, given the k first labels on the path from v's parent to the root (its *k-label-history*)



2nd-order label entropy:

- 2-history *ab*: 5 nodes total, 3 nodes labeled *a*, 2 nodes labeled *c*
- 2-history *ba*: 3 nodes total, 2 nodes labeled *a*, 1 node labeled *c*
- 2-history *aa*: 1 node labeled *b*

Label Entropy H_k^ℓ



2nd-order label entropy:

- 2-history *ab*: 5 nodes total, 3 nodes labeled *a*, 2 nodes labeled *c*
- 2-history *ba*: 3 nodes total, 2 nodes labeled *a*, 1 node labeled *c*
- 2-history *aa*: 1 node labeled *b*

$$H_2^\ell(t) = 3\log_2\left(\frac{5}{3}\right) + 2\log_2\left(\frac{5}{2}\right) + 2\log_2\left(\frac{3}{2}\right) + 1\log_2\left(\frac{3}{1}\right) + 1\log_2\left(\frac{1}{1}\right) \approx 7,6096$$

Label Entropy H_k^ℓ

Theorem (Ganczorz, 2020)

Let t be a labeled tree, then t can be represented in

$$H^{\deg}(t) + H_k^\ell(t)$$

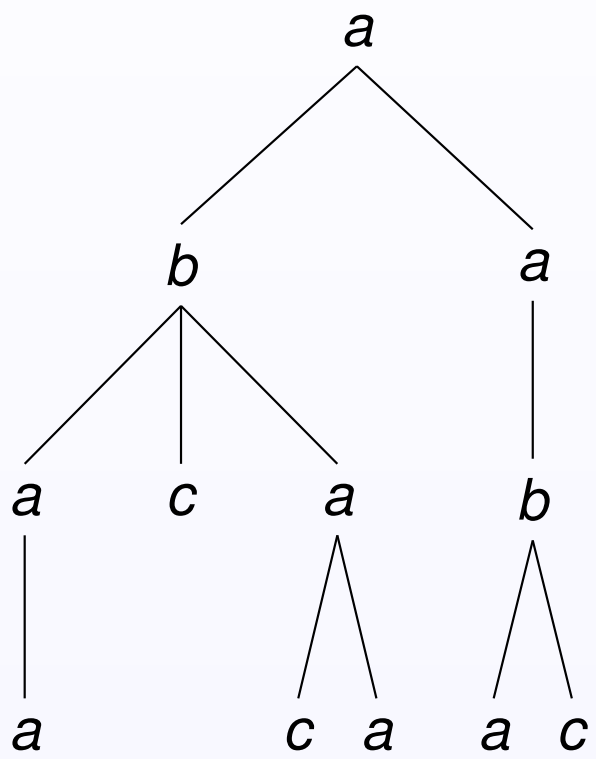
(plus lower-order term) many bits.

Degree-Label Entropy $H_k^{\deg,\ell}$

- Defined by Ganczorz in 2020
- Expected uncertainty about the label of a node v, given
  - (i) its k-label-history and
  - (ii) the degree of the node

Degree-Label Entropy $H_k^{\deg,\ell}$

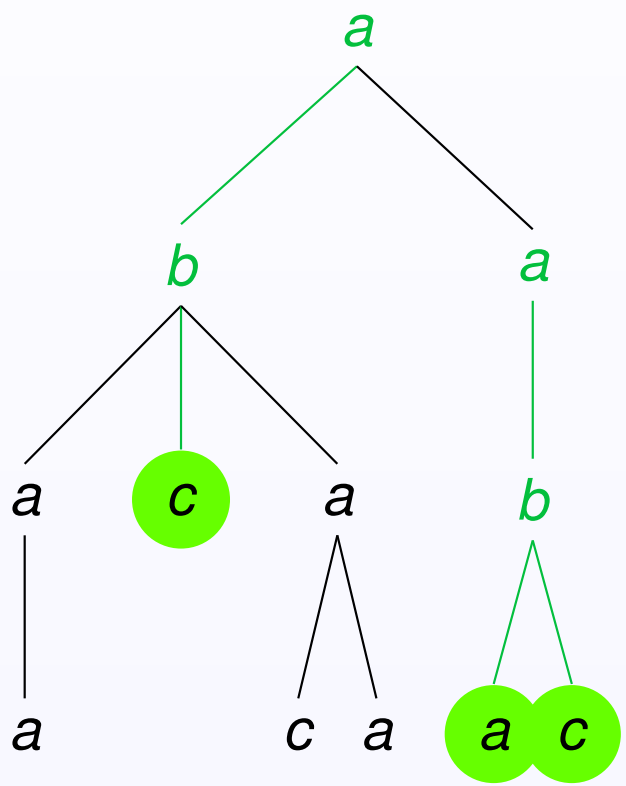
- Defined by Ganczorz in 2020
- Expected uncertainty about the label of a node v, given
  - (i) its k-label-history and
  - (ii) the degree of the node



2nd-order degree-label entropy:

Degree-Label Entropy $H_k^{\deg,\ell}$

- Defined by Ganczorz in 2020
- Expected uncertainty about the label of a node v, given
  - (i) its k-label-history and
  - (ii) the degree of the node

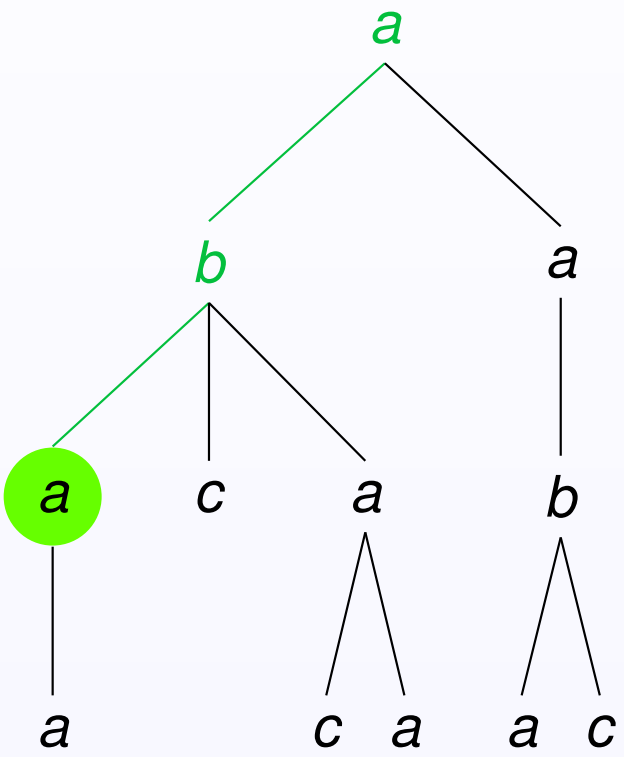


2nd-order degree-label entropy:

- 2-history *ab*, degree 0: 2 nodes labeled *c*, 1 node labeled *a*

Degree-Label Entropy $H_k^{\deg,\ell}$

- Defined by Ganczorz in 2020
- Expected uncertainty about the label of a node v, given
  - (i) its k-label-history and
  - (ii) the degree of the node



2nd-order degree-label entropy:

- 2-history *ab*, degree 0: 2 nodes labeled *c*, 1 node labeled *a*
- 2-history *ab*, degree 1: 1 node labeled *a*
- (...)

Degree-Label Entropy $H_k^{\deg,\ell}$

Theorem (Ganczorz, 2020)

Let t be a labeled tree, then t can be represented in

$$H_k^{\deg,\ell}(t) + H^{\deg}(t)$$

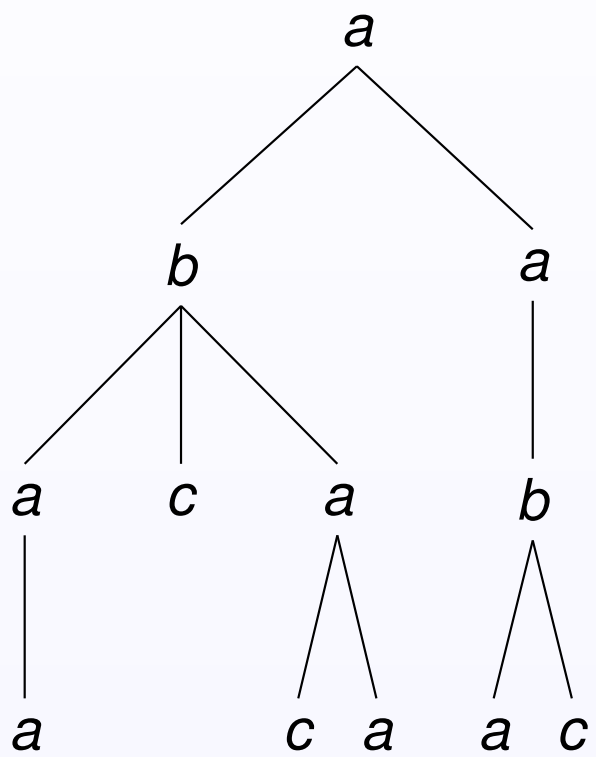
(plus lower-order term) many bits.

Label-Degree Entropy $H_k^{\ell,\deg}$

- Defined by Ganczorz in 2020
- Expected uncertainty about the degree of a node v, given
  - (i) its k-label-history and
  - (ii) the label of the node

Label-Degree Entropy $H_k^{\ell,\deg}$

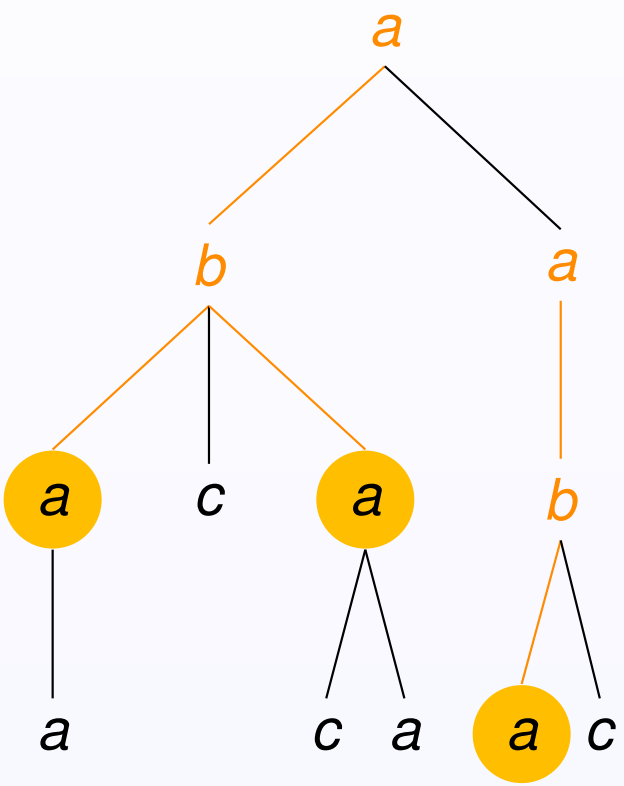
- Defined by Ganczorz in 2020
- Expected uncertainty about the degree of a node v, given
  - (i) its k-label-history and
  - (ii) the label of the node



2nd-order label-degree entropy:

Label-Degree Entropy $H_k^{\ell,\deg}$

- Defined by Ganczorz in 2020
- Expected uncertainty about the degree of a node v, given
  - (i) its k-label-history and
  - (ii) the label of the node



2nd-order label-degree entropy:

- 2-history *ab* and label *a*: 1 node of degree 2, 1 node of degree 1, 1 node of degree 0
- (...)

Label-Degree Entropy $H_k^{\ell,\deg}$

Theorem (Ganczorz, 2020)

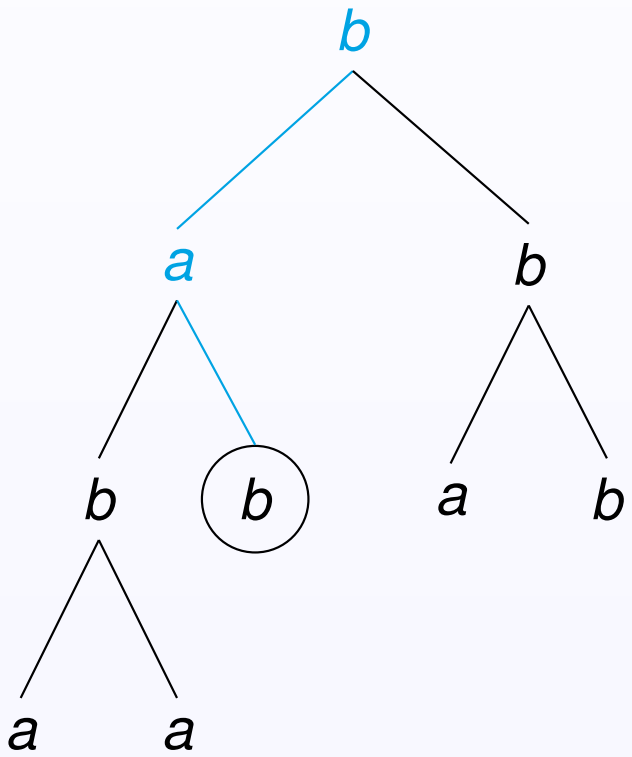
Let t be a labeled tree, then t can be represented in

$$H_k^{\ell,\deg}(t) + H_k^{\ell}(t)$$

(plus lower-order term) many bits.

Label-Shape Entropy $H_k^{\ell s}$ for binary trees

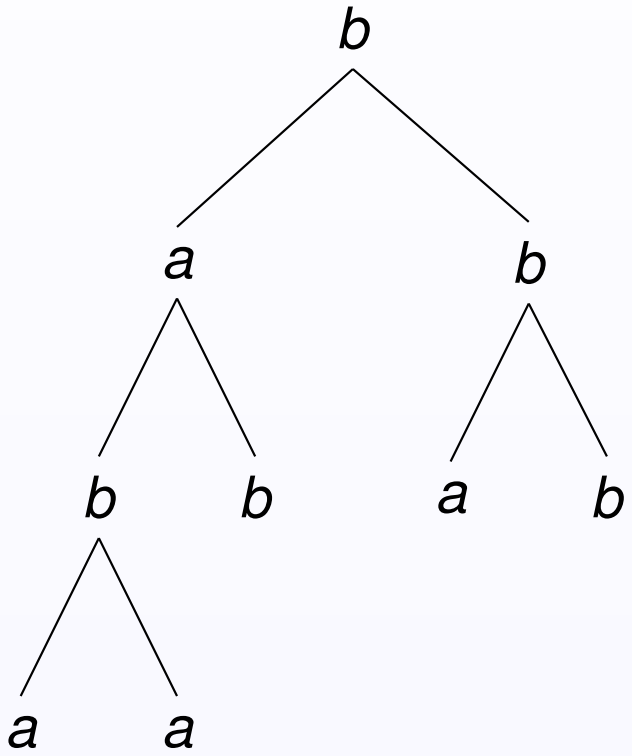
- Defined by Hucke et al. in 2019
- Binary trees: each node of the tree has either 0 or 2 descendants
- k-Label-Shape-History of a node: String consisting of the last k node labels and directions (0 for left, 1 for right) on the path from the root to v's parent



1-Label-Shape-History: a1
2-Label-Shape-History: b0a1

Label-Shape Entropy $H_k^{\ell s}$ for binary trees

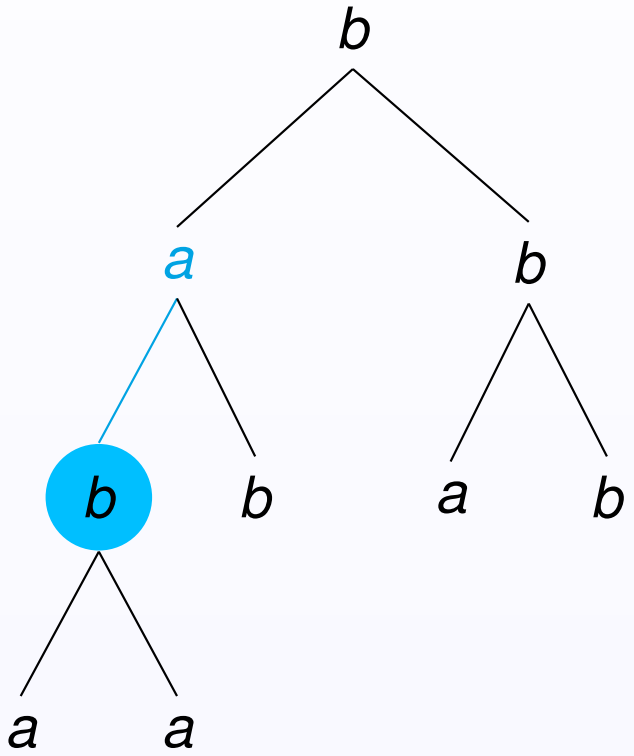
- k-th order Label-Shape-Entropy is the expected uncertainty of $(\text{Label}(v), \text{Degree}(v))$ of a node v, given its k-label-shape-history:



1st order label shape entropy:

Label-Shape Entropy $H_k^{\ell s}$ for binary trees

- k-th order Label-Shape-Entropy is the expected uncertainty of $(\text{Label}(v), \text{Degree}(v))$ of a node v, given its k-label-shape-history:

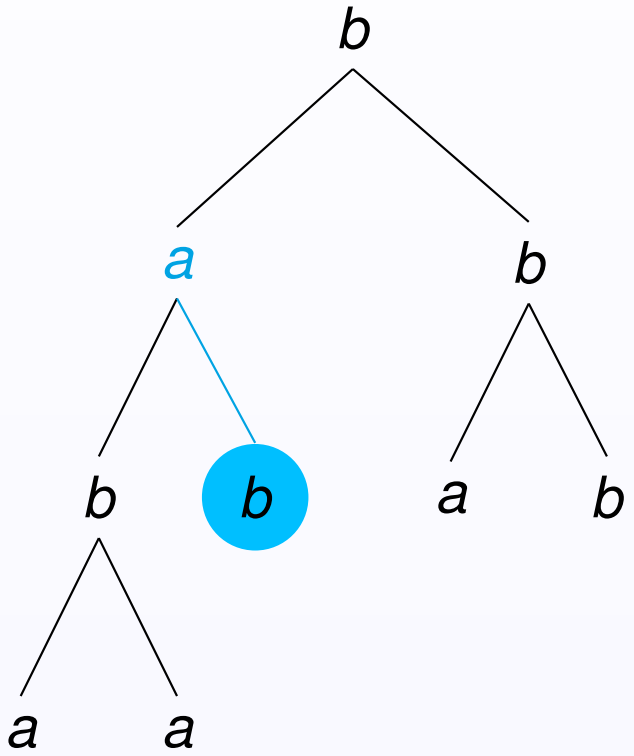


1st order label shape entropy:

- 1-history a0: 1 binary node labeled b

Label-Shape Entropy $H_k^{\ell s}$ for binary trees

- k-th order Label-Shape-Entropy is the expected uncertainty of $(\text{Label}(v), \text{Degree}(v))$ of a node v, given its k-label-shape-history:

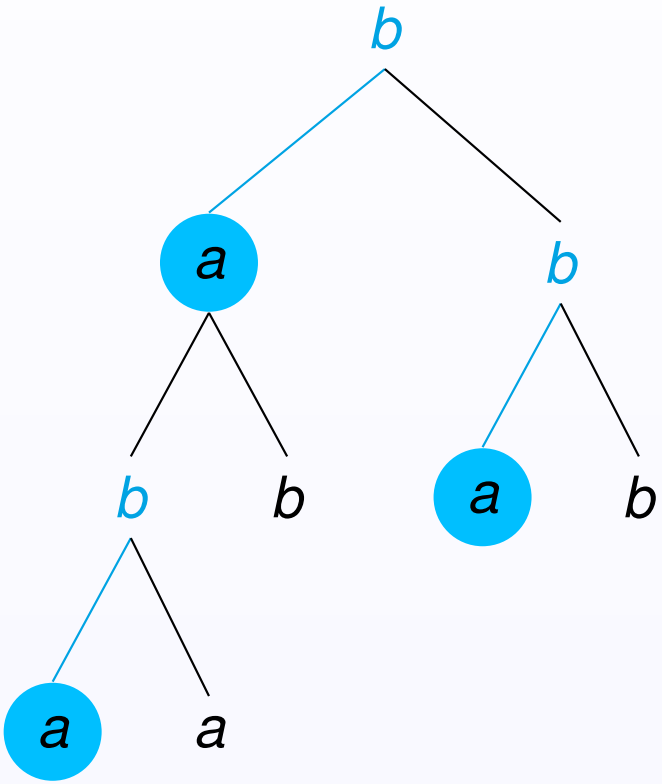


1st order label shape entropy:

- 1-history a0: 1 binary node labeled b
- 1-history a1: 1 leaf labeled b

Label-Shape Entropy $H_k^{\ell s}$ for binary trees

- k-th order Label-Shape-Entropy is the expected uncertainty of $(\text{Label}(v), \text{Degree}(v))$ of a node v, given its k-label-shape-history:

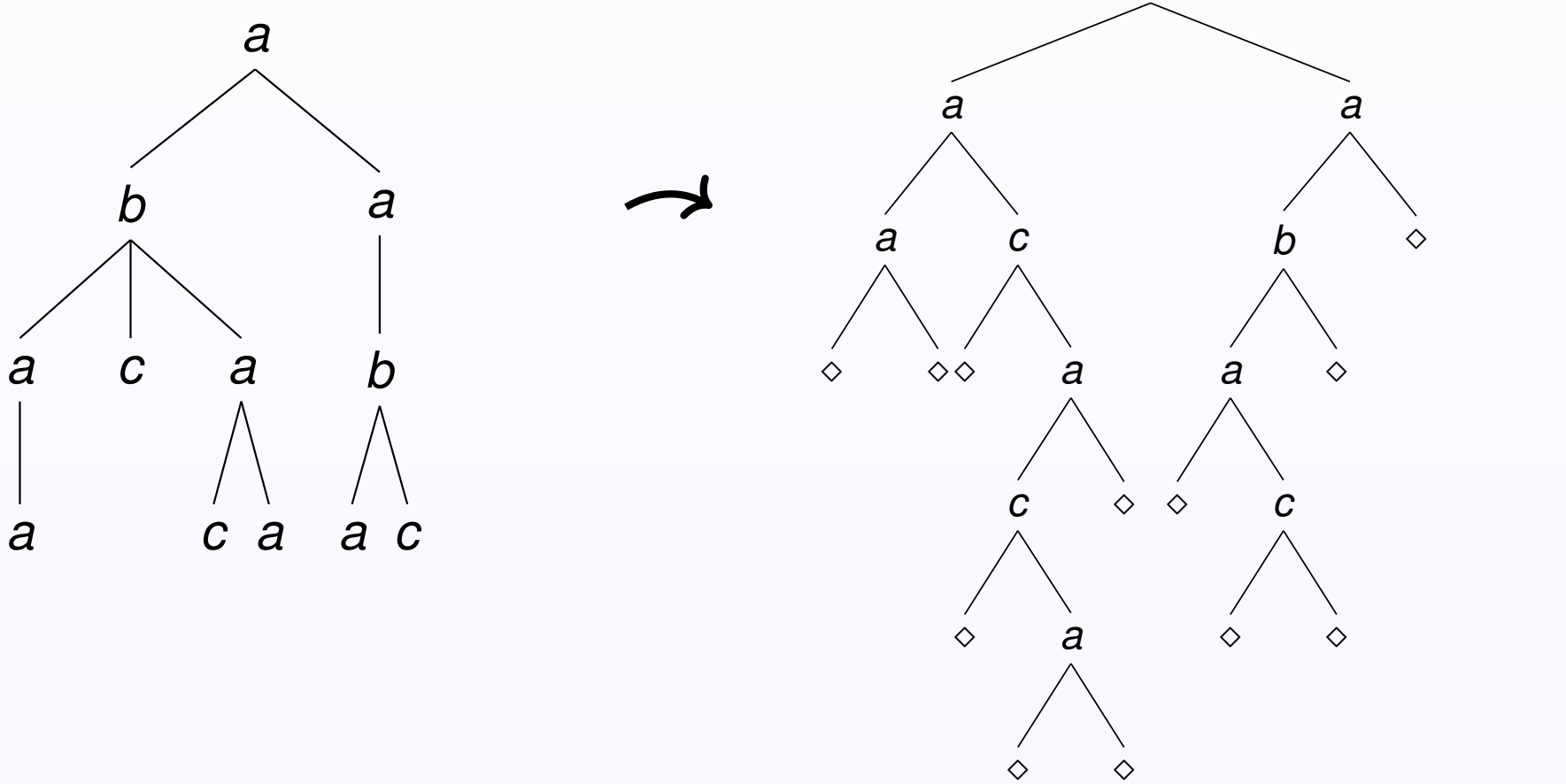


1st order label shape entropy:

- 1-history a0: 1 binary node labeled b
- 1-history a1: 1 leaf labeled b
- 1-history b0: 1 binary node labeled a, 2 leaves labeled a
- (...)

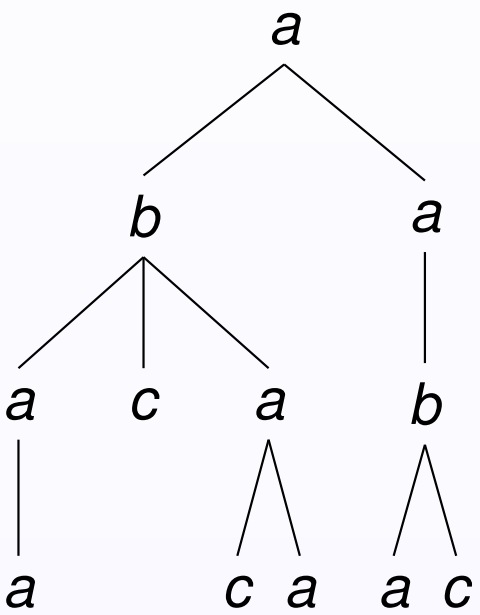
Label-Shape Entropy $H_k^{\ell s}$ for unranked trees

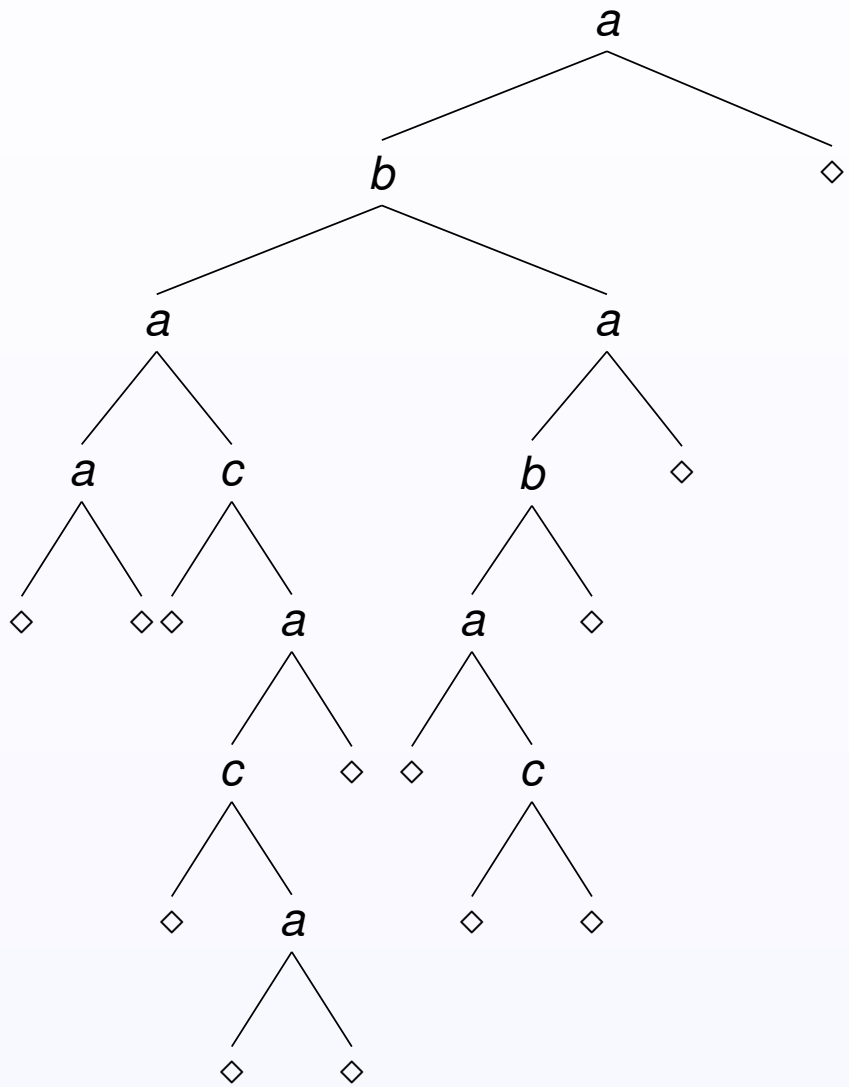
- Unranked trees can be transformed into binary trees via first-child next-sibling encoding (fcns)



Label-Shape Entropy $H_k^{\ell s}$ for unranked trees

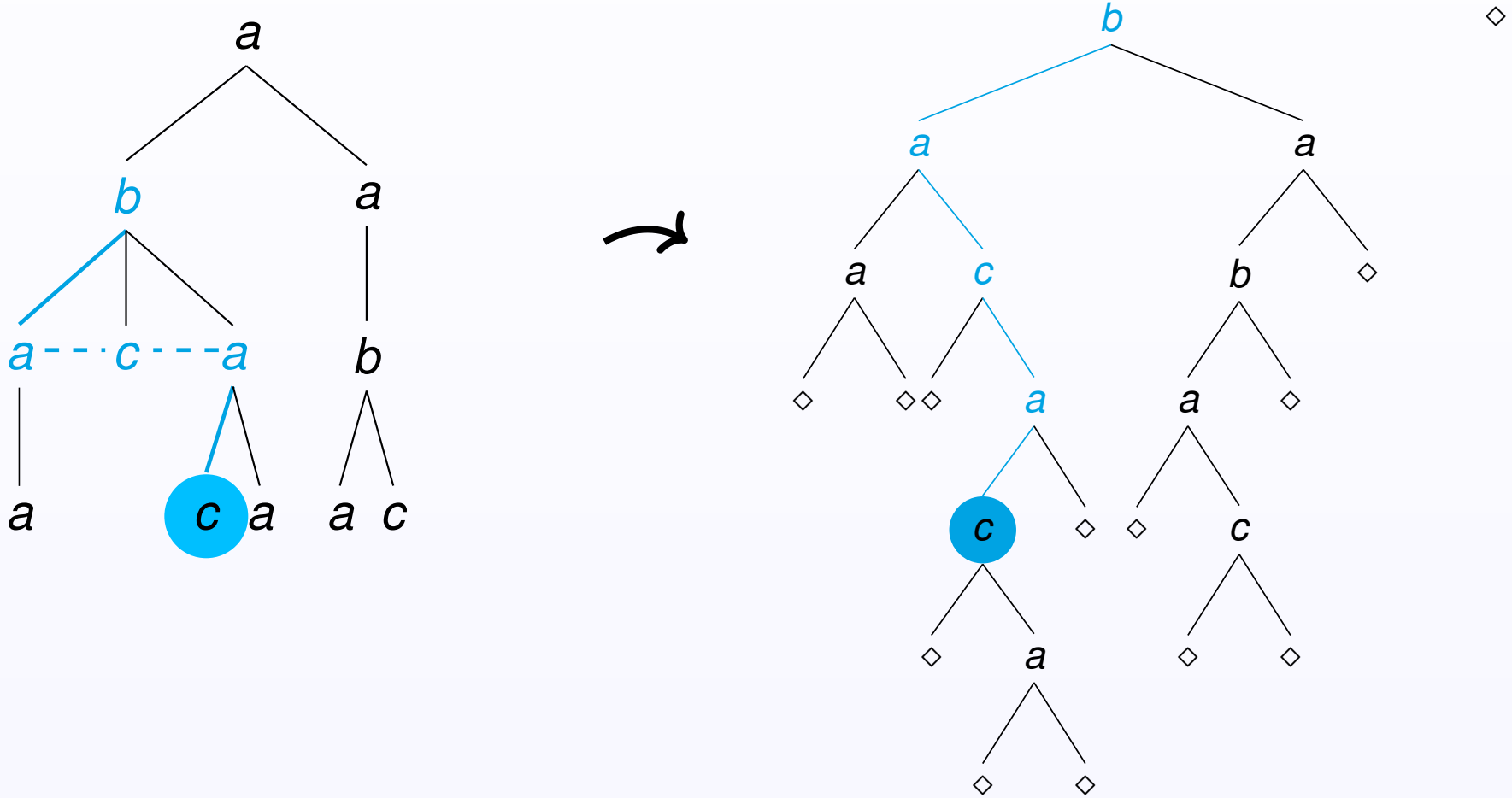
- We define $H_k^{\ell s}(t) = H_k^{\ell s}(\mathrm{fcns}(t))$ for unranked trees t





Label-Shape Entropy $H_k^{\ell s}$ for unranked trees

- We define $H_k^{\ell s}(t) = H_k^{\ell s}(\mathrm{fcns}(t))$ for unranked trees t



Label-Shape Entropy $H_k^{\ell s}$

Theorem (Hucke, Lohrey, S.B., 2019)

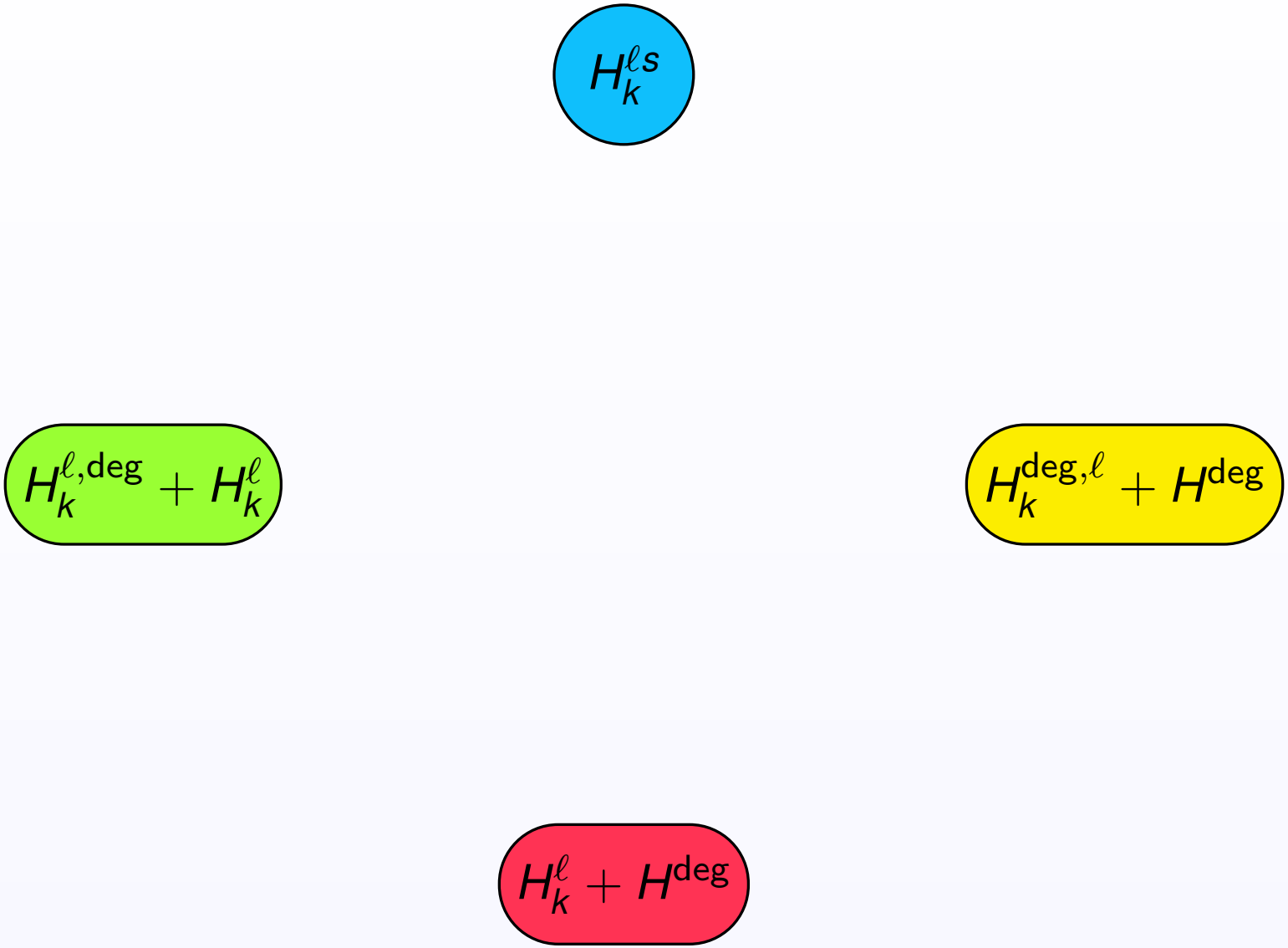
Every labeled tree t can be represented in

$$H_k^{\ell s}(t)$$

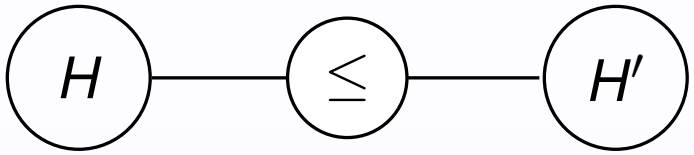
(plus lower-order term) many bits.

- Representation is based on Tree-Straight-Line Programms (TSLPs)

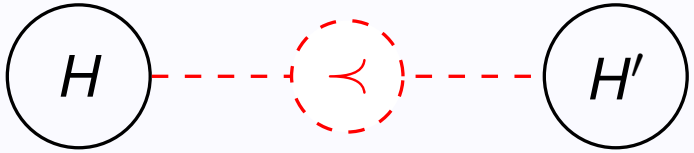
Entropy Bounds for Tree Compressors



Entropy Bounds for Tree Compressors



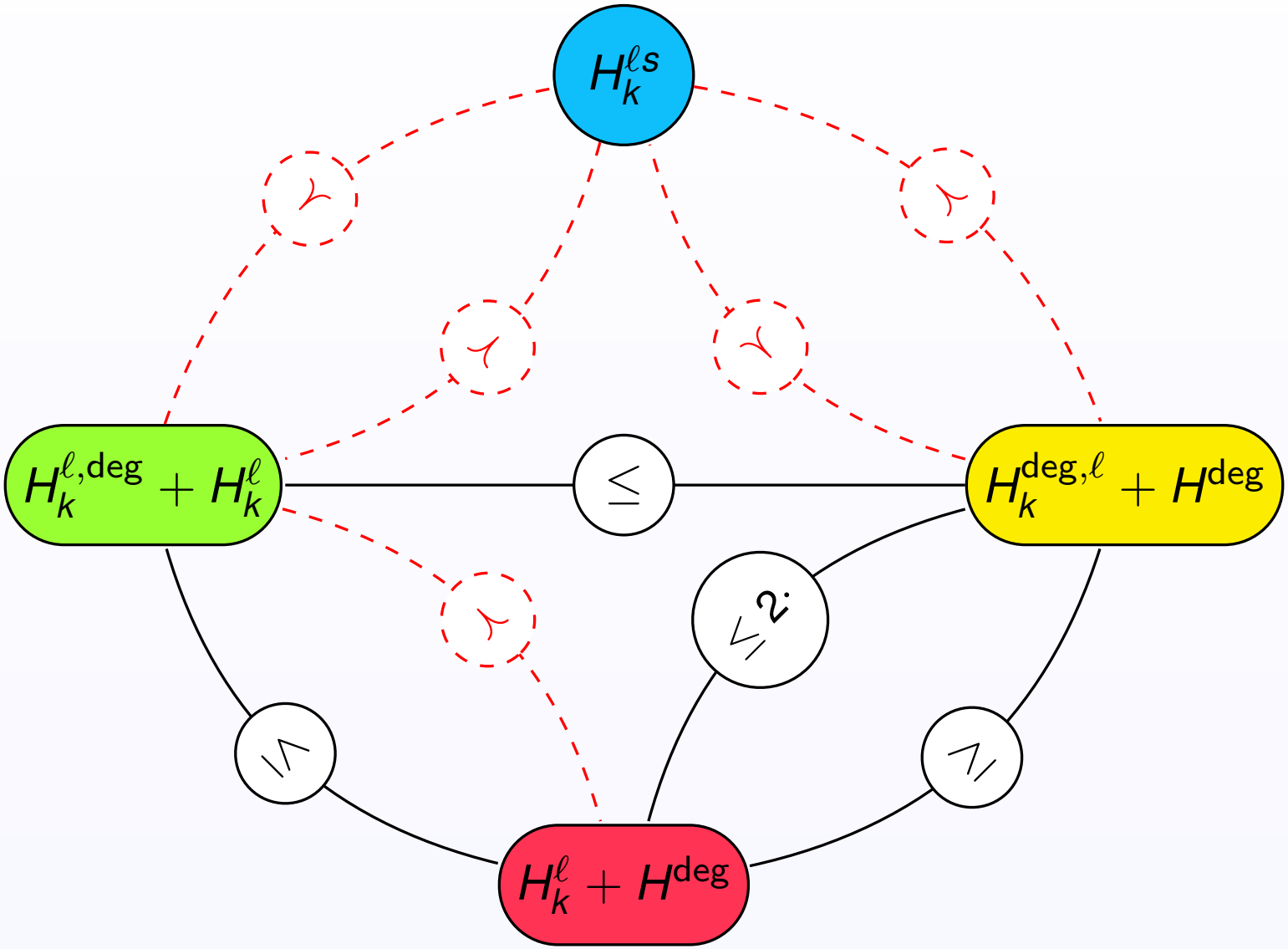
For all trees t, we have $H(t) \leq H'(t)$.



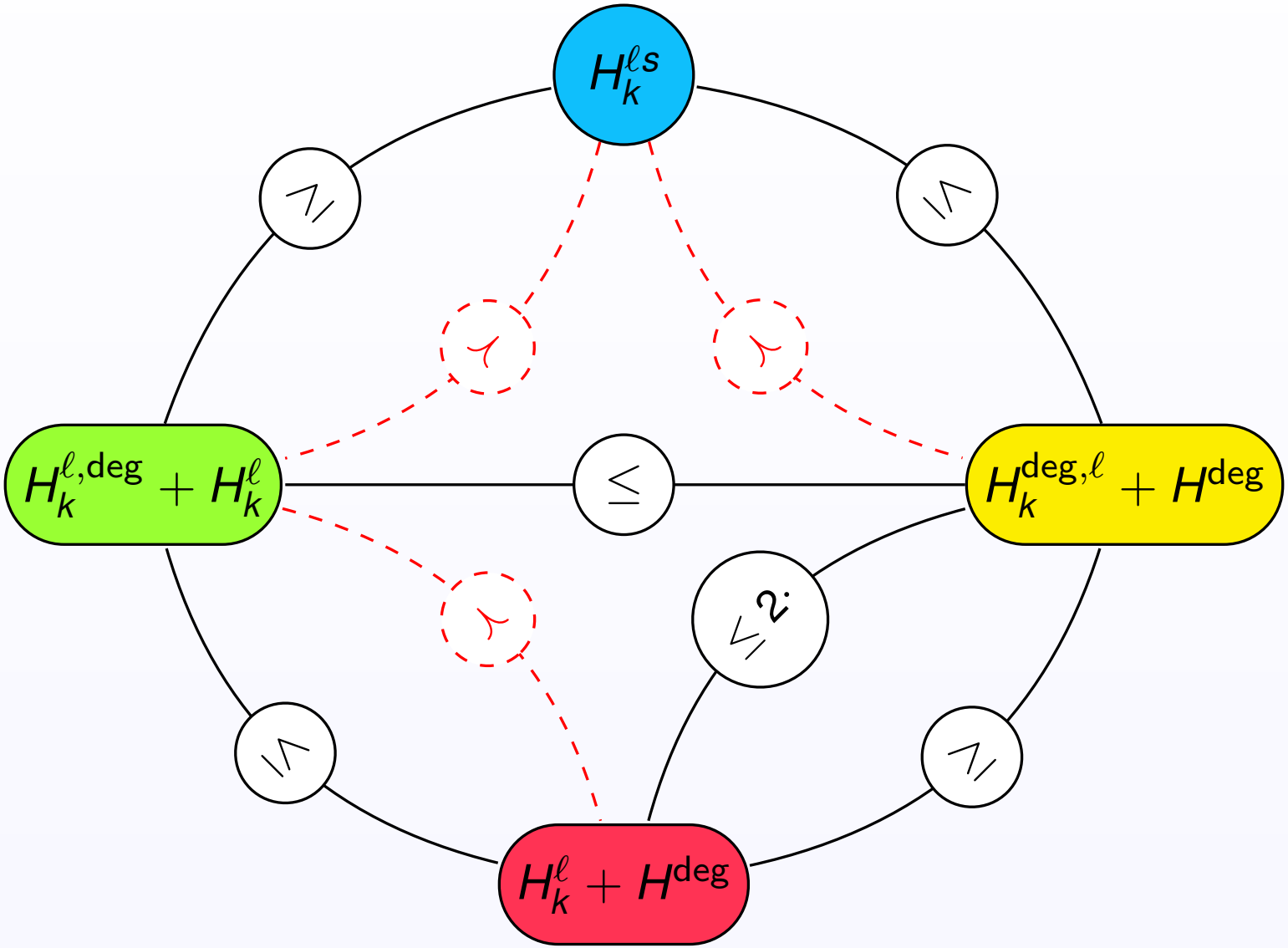
There exists a family of trees $(t_n)_{n \in \mathbb{N}}$, such that

- $|t_n| \in \Theta(n)$ and
- $H(t_n) \in o(H'(t_n))$.

Entropy Bounds for Tree Compressors: Labeled Unranked Trees



Entropy Bounds for Tree Compressors: Labeled Binary Trees



Experimental Comparison

XML	k	$H_k^{\ell s}$	$H^{\deg} + H_k^{\ell}$	$H_k^{\ell} + H_k^{\ell,\deg}$	$H^{\deg} + H_k^{\deg,\ell}$
DBLP*	0	18 727 523.4	14 576 781.0	12 967 501.2	12 967 501.2
	1	2 607 784.7	12 137 042.6	10 527 690.4	12 076 935.4
	2	2 076 410.5	12 136 974.7	10 527 595.9	12 076 845.7
	4	1 951 141.6	12 136 966.2	10 527 586.3	12 076 836.8

XML	k	$H_k^{\ell s}/H_k^{\ell s}$	$(H^{\deg} + H_k^{\ell})/H_k^{\ell s}$	$(H_k^{\ell} + H_k^{\ell,\deg})/H_k^{\ell s}$	$(H^{\deg} + H_k^{\deg,\ell})/H_k^{\ell s}$
DBLP*	0	1	0.79	0.69	0.69
	1	1	4.65	4.04	4.63
	2	1	5.84	5.07	5.81
	4	1	6.22	5.39	6.19

* from http://xmlcompbench.sourceforge.net

Experimental Comparison

XML	k	$H_k^{\ell s}/H_k^{\ell s}$	$(H^{\deg}+H_k^{\ell})/H_k^{\ell s}$	$(H_k^{\ell}+H_k^{\ell,\deg})/H_k^{\ell s}$	$(H^{\deg}+H_k^{\deg,\ell})/H_k^{\ell s}$
DBLP	0	1	0.79	0.69	0.69
	1	1	4.65	4.04	4.63
	2	1	5.84	5.07	5.81
	4	1	6.22	5.39	6.19
BaseBall	0	1	0.75	0.72	0.72
	1	1	22.95	21.73	22.89
	2	1	54.53	51.63	54.38
	4	1	101.53	96.13	101.26
Treebank	0	1	0.97	0.8	0.8
	1	1	1.64	1.26	1.32
	2	1	2.12	1.61	1.70
	4	1	2.49	1.87	1.99
Average*	0	1	0.88	0.71	0.71
	1	1	13.34	6.24	12.72
	2	1	21.40	12.30	19.89
	4	1	25.90	16.37	24.27

* Average over 13 files from http://xmlcompbench.sourceforge.net

Thank you for your attention!

References


Paolo Ferragina, Fabrizio Luccio, Giovanni Manzini, and S. Muthukrishnan.
Structuring labeled trees for optimal succinctness, and beyond.
In *Proc. FOCS 2005*, pages 184–196. IEEE Computer Society, 2005.


Danny Hucke, Markus Lohrey, and Louisa Seelbach Benkner.
Entropy bounds for grammar-based tree compressors.
In *Proc. ISIT 2019*, pages 1687–1691. IEEE, 2019.


Michal Ganczorz.
Using statistical encoding to achieve tree succinctness never seen before.
In *Proc. STACS 2020*, volume 154 of *LIPIcs*, pages 22:1–22:29.


Jesper Jansson, Kunihiko Sadakane, and Wing-Kin Sung.
Ultra-succinct representation of ordered trees with applications.
Journal of Computer and System Sciences, 78(2):619–631, 2012.