High order entropy

Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Department of Computer Science

High-order entropy

Zero order empirical entropy seems insufficient when context matters

Bigram frequency per 40,000 words

th	1.52	**en**	0.55	**ng**	0.18
he	1.28	**ed**	0.53	**of**	0.16
in	0.94	**to**	0.52	**al**	0.09
er	0.94	**it**	0.50	**de**	0.09
an	0.82	**ou**	0.50	**se**	0.08
re	0.68	**ea**	0.47	**le**	0.08
nd	0.63	**hi**	0.46	**sa**	0.06
at	0.59	**is**	0.46	**si**	0.05
on	0.57	**or**	0.43	**ar**	0.04
nt	0.56	**ti**	0.34	**ve**	0.04
ha	0.56	**as**	0.33	**ra**	0.04
es	0.56	**te**	0.27	**ld**	0.02
st	0.55	**et**	0.19	**ur**	0.02

http://math.cornell.edu/~mec/2003-2004/cryptography/subs/digraphs.html

High-order entropy

Can compress better if we consider ***context***

Let C change depending on surrounding symbols

gtgtatcggagcgctctgcgttatcgatcgcgatctggt

$C_{\mathtt{tcg}}(\mathtt{a})$ $C_{\mathtt{gcg}}(\mathtt{a})$ $C_{\mathtt{tcg}}(\mathtt{a})$ $C_{\mathtt{tcg}}(\mathtt{a})$ $C_{\mathtt{tct}}(\mathtt{a})$

For k symbols of context, we have codes $C_i \in \{C_{\Sigma^k}\}$

High-order entropy

Could consider context to the right →

gtgtatcggagcgctctgcgttatcgatcgcgatctggt

$C_{\mathtt{tcg}}(\mathtt{a})$ $C_{\mathtt{gcg}}(\mathtt{a})$ $C_{\mathtt{tcg}}(\mathtt{a})$ $C_{\mathtt{tcg}}(\mathtt{a})$ $C_{\mathtt{tct}}(\mathtt{a})$

Or context to the left ←

gtgtatcggagcgctctgcgttatcgatcgcgatctggt

$C_{\mathtt{tct}}(\mathtt{a})$ $C_{\mathtt{cgg}}(\mathtt{a})$ $C_{\mathtt{gtt}}(\mathtt{a})$ $C_{\mathtt{tcg}}(\mathtt{a})$ $C_{\mathtt{gcg}}(\mathtt{a})$

High-order entropy

How should we build each code $\{C_{\Sigma^k}\}$?

Same as before, but with frequencies *conditioned on context*

E.g. $C_{\mathtt{gca}}$ is built considering the number of times each symbol occurs just after gca

High-order entropy

abracadabraabracadabra

Let S_a be the string we get by concatenating characters just after the a's

$S_a =$

High-order entropy

abracadabraabracadabra

Let $S_{\mathtt{a}}$ be the string we get by concatenating characters just after the a's

$S_{\mathtt{a}} = \mathtt{bcdbabcdb}$

Build code $C_{\mathtt{a}}$ using frequencies in S_a:

High-order entropy

abracadabraabracadabra

Let S_a be the string we get by concatenating characters just after the a's

S_a = bcdbabcdb

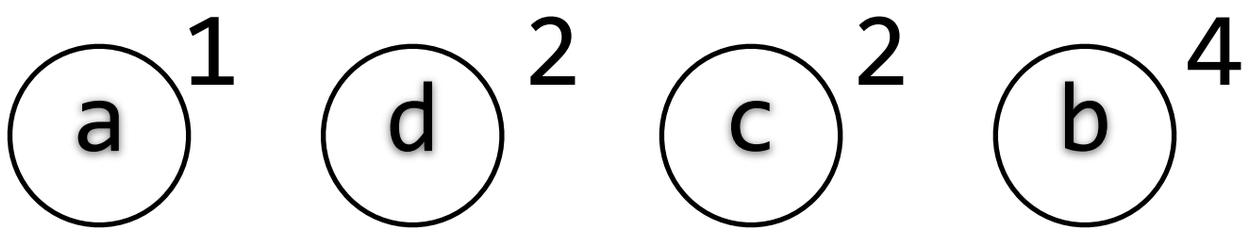
Build code C_a using frequencies in S_a:

{a : 1, b : 4, c : 2, d : 2, r : 0}

(r won't get a code)

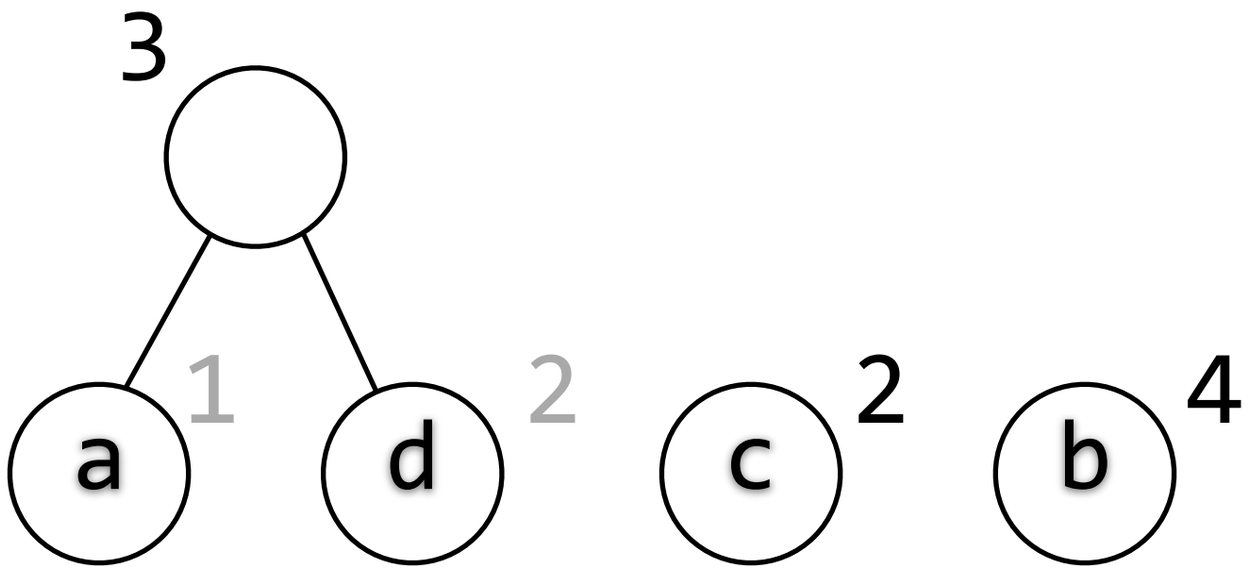
High-order entropy

$$\{\mathtt{a} : 1,\ \mathtt{b} : 4,\ \mathtt{c} : 2,\ \mathtt{d} : 2,\ \mathtt{r} : 0\}$$



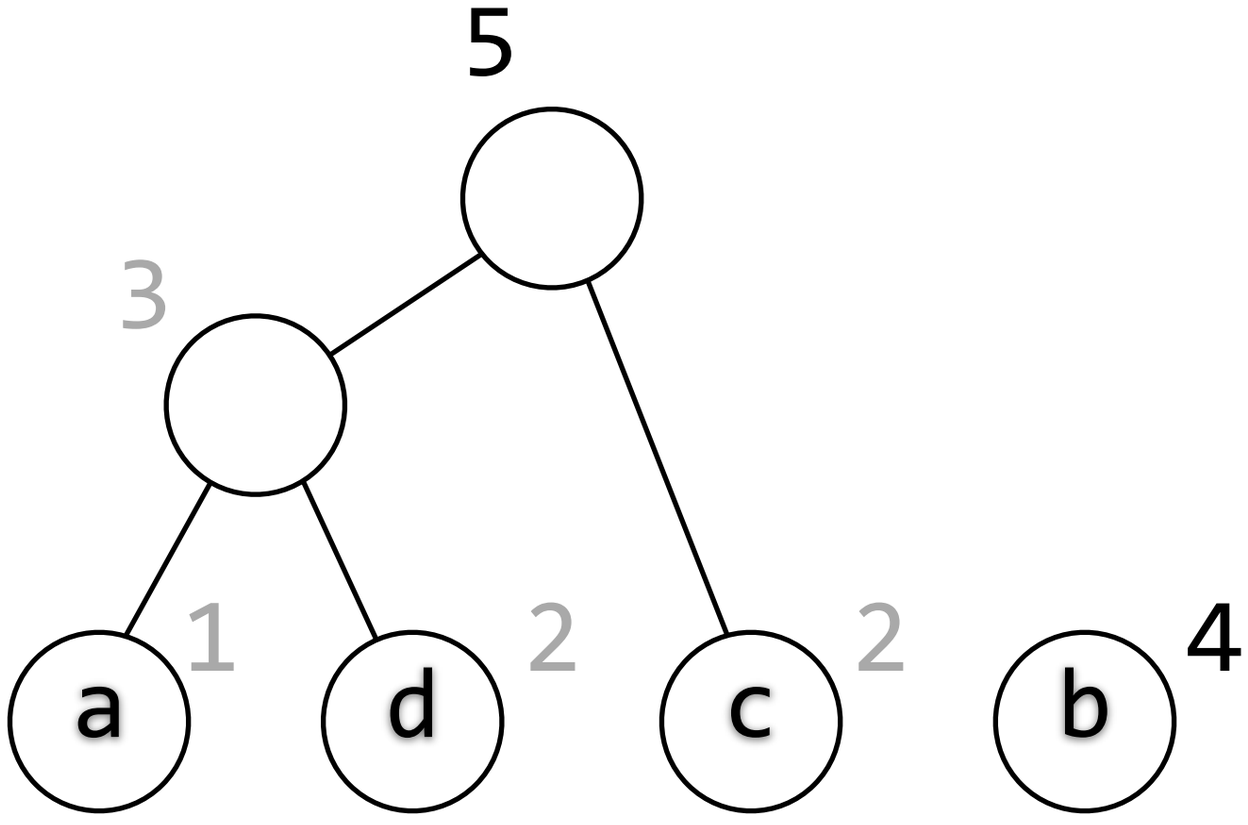
High-order entropy

$$\{\mathtt{a} : 1,\ \mathtt{b} : 4,\ \mathtt{c} : 2,\ \mathtt{d} : 2,\ \mathtt{r} : 0\}$$



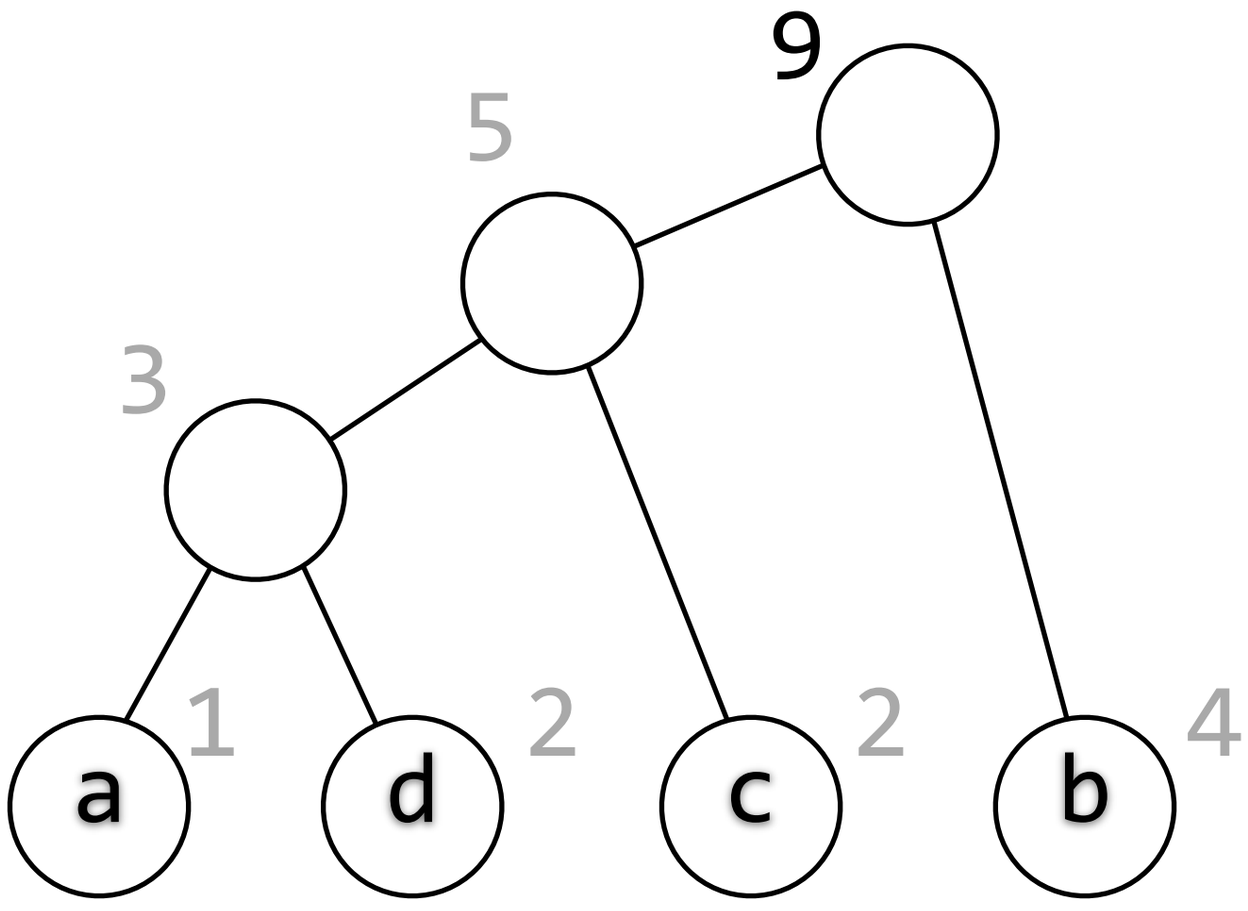
High-order entropy

$$\{\mathtt{a} : 1,\ \mathtt{b} : 4,\ \mathtt{c} : 2,\ \mathtt{d} : 2,\ \mathtt{r} : 0\}$$



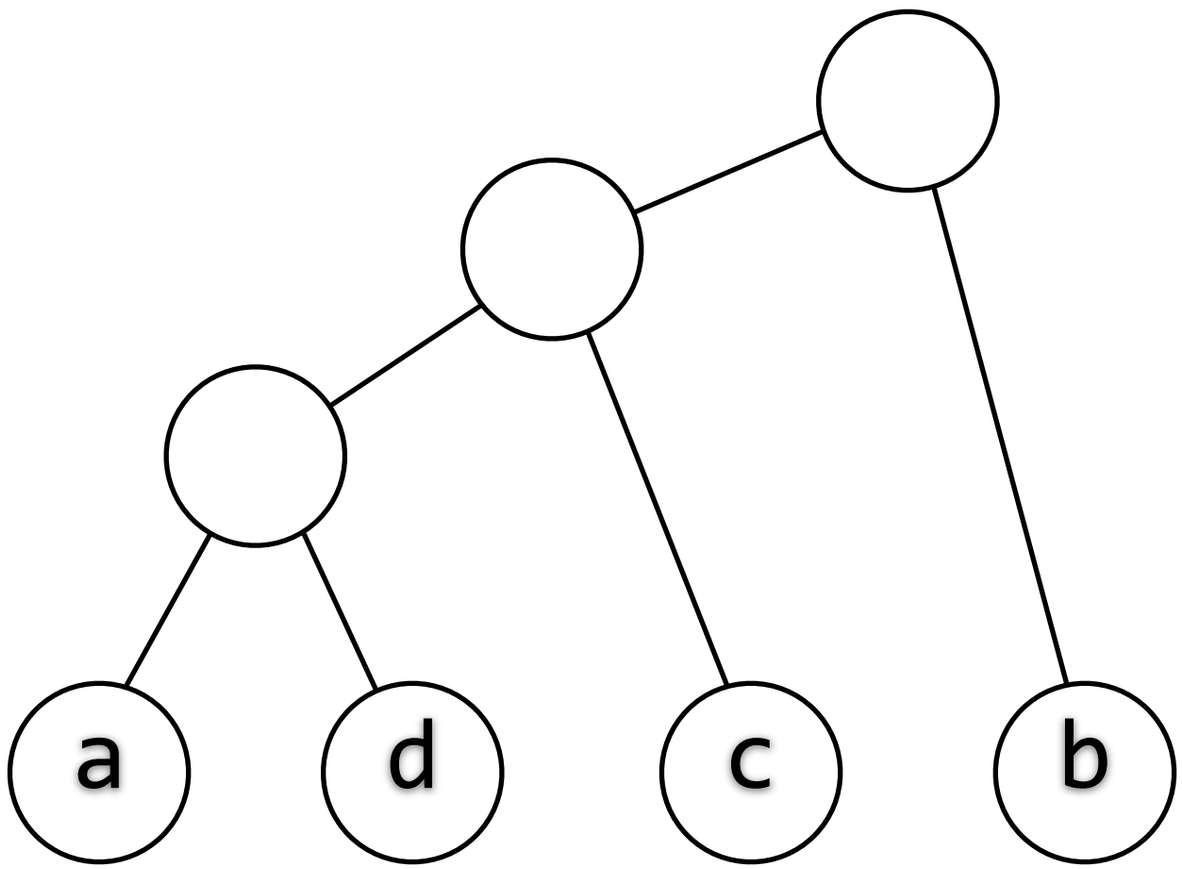
High-order entropy

$$\{\mathtt{a}:1,\ \mathtt{b}:4,\ \mathtt{c}:2,\ \mathtt{d}:2,\ \mathtt{r}:0\}$$



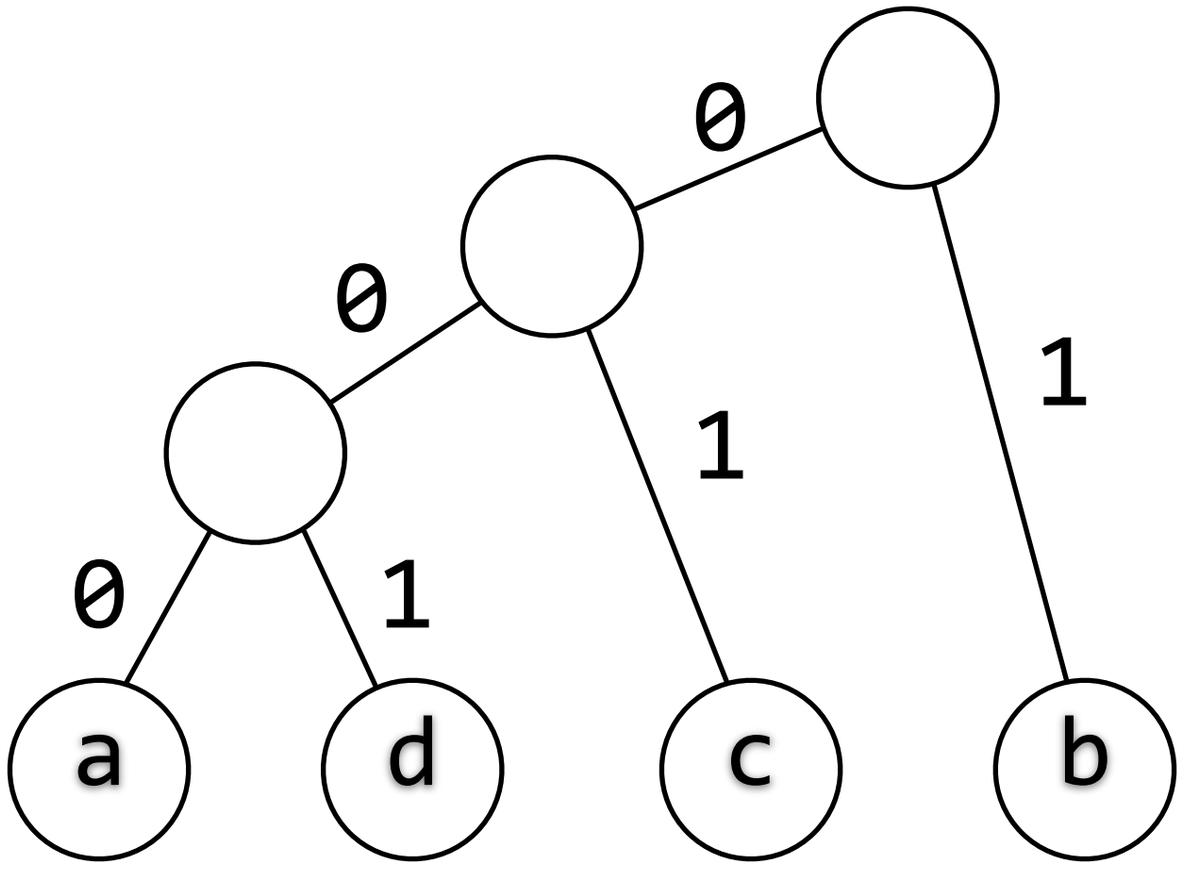
High-order entropy

$$\{\mathtt{a}:1,\ \mathtt{b}:4,\ \mathtt{c}:2,\ \mathtt{d}:2,\ \mathtt{r}:0\}$$



High-order entropy

$$\{\mathtt{a}:1,\ \mathtt{b}:4,\ \mathtt{c}:2,\ \mathtt{d}:2,\ \mathtt{r}:0\}$$



$$C_{\mathtt{a}}(\mathtt{a}) = 000$$

$$C_{\mathtt{a}}(\mathtt{d}) = 001$$

$$C_{\mathtt{a}}(\mathtt{c}) = 01$$

$$C_{\mathtt{a}}(\mathtt{b}) = 1$$

High-order entropy

`mississippimississippi`

$S_{\mathtt{i}} = \mathtt{sspmssp}$

$\{\mathtt{s}:4,\ \mathtt{p}:2,\ \mathtt{m}:1\}$

$C_{\mathtt{i}}(\mathtt{p}) = 00$

$C_{\mathtt{i}}(\mathtt{m}) = 01$

$C_{\mathtt{i}}(\mathtt{s}) = 1$

$S_{\mathtt{m}} = \mathtt{ii} \quad \{\mathtt{i}:2\}$

(no code)

$S_{\mathtt{p}} = \mathtt{pipi} \qquad C_{\mathtt{p}}(\mathtt{p}) = 0$

$\{\mathtt{p}:2,\ \mathtt{i}:2\} \qquad C_{\mathtt{p}}(\mathtt{i}) = 1$

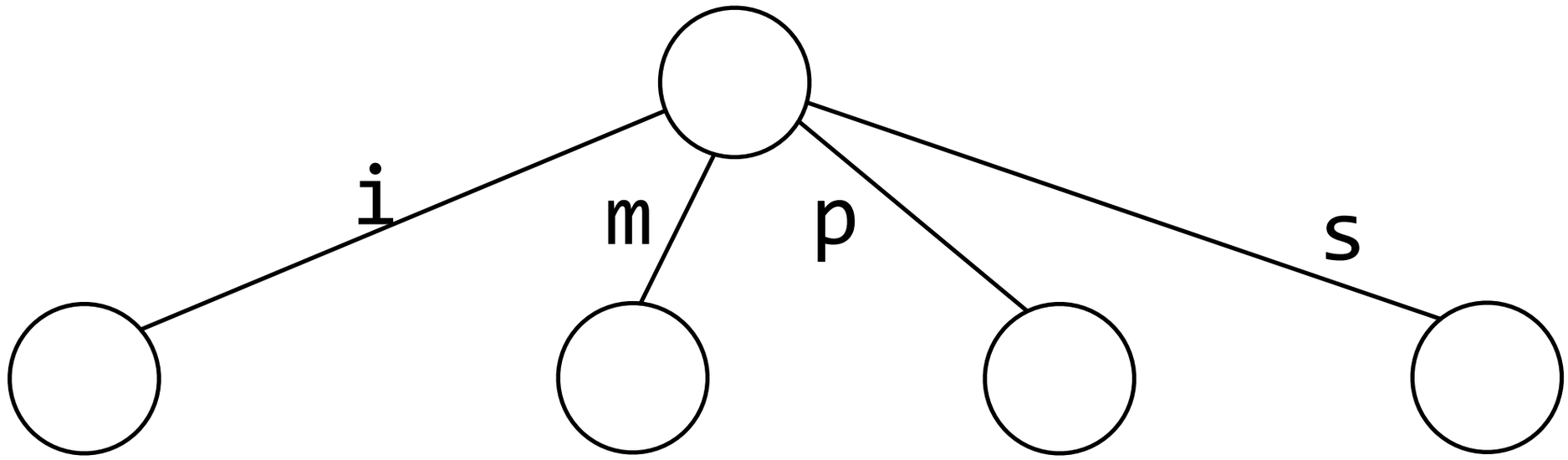
$S_{\mathtt{s}} = \mathtt{sisisisi}$

$\{\mathtt{s}:4,\ \mathtt{i}:4\} \qquad C_{\mathtt{s}}(\mathtt{s}) = 0$

$C_{\mathtt{s}}(\mathtt{i}) = 1$

High-order entropy

mississippimississippi



$C_{\mathtt{i}}(\mathtt{p}) = 00$

$C_{\mathtt{i}}(\mathtt{m}) = 01$

$C_{\mathtt{i}}(\mathtt{s}) = 1$

i

(no code)

$C_{\mathtt{p}}(\mathtt{p}) = 0$

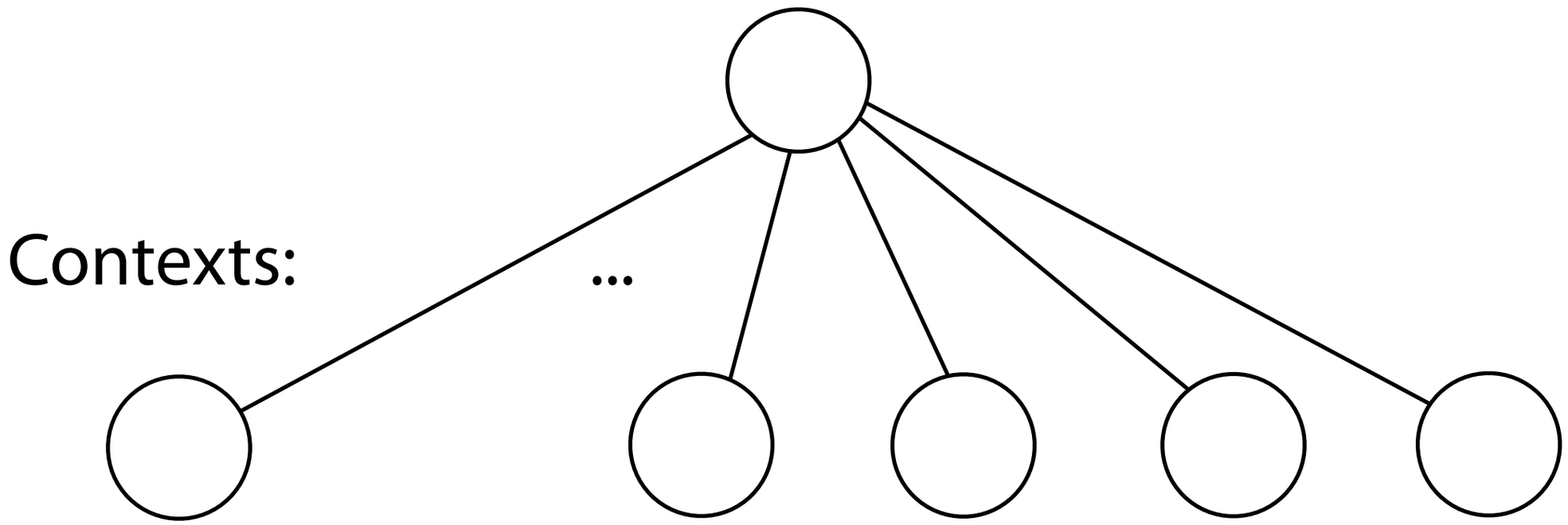
$C_{\mathtt{p}}(\mathtt{i}) = 1$

$C_{\mathtt{s}}(\mathtt{s}) = 0$

$C_{\mathtt{s}}(\mathtt{i}) = 1$

High-order entropy

Def'n of ***high-order empirical entropy*** H_k is similarly hierarchical



Codes achieving near-H_0 given context

High-order entropy

H_k of length-n string S is a weighted sum over all contexts of the zero order empirical entropy of symbols having that context

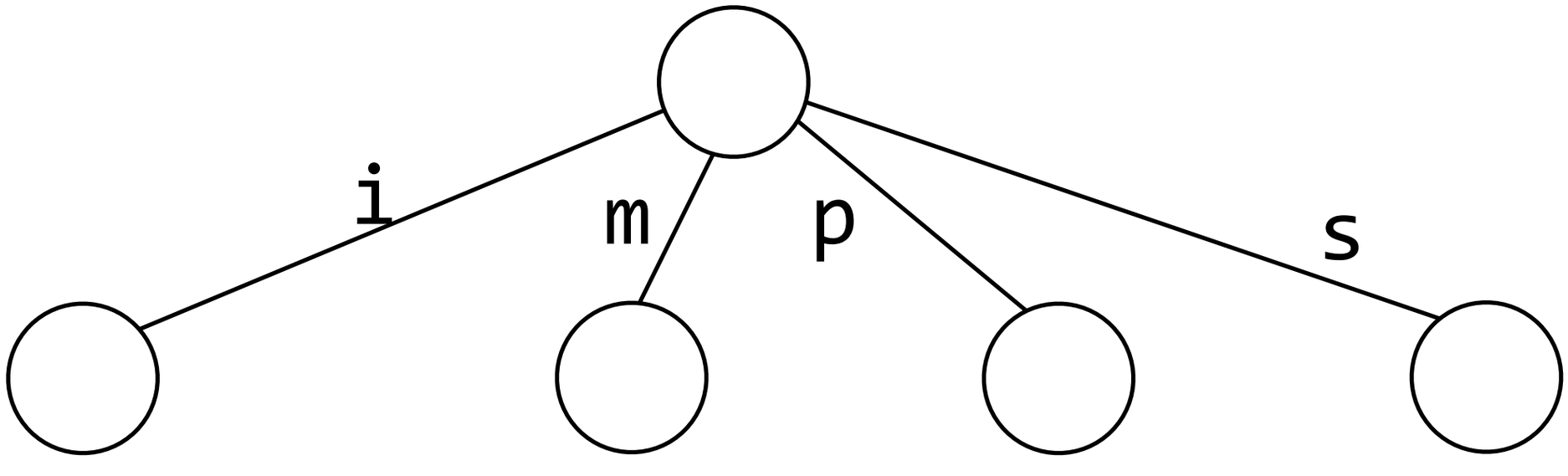
High-order entropy

H_k of length-n string S is a weighted sum over all contexts of the zero order empirical entropy of symbols having that context

$$H_k(S) = \sum_{t \in \Sigma^k} \frac{|S_t|}{n} \cdot H_0(S_t) \qquad \text{for } k > 0$$

S_t is the concatenation of symbols having context t

High-order entropy



$C_{\texttt{i}}(\texttt{p}) = 00$

$C_{\texttt{i}}(\texttt{m}) = 01$

$C_{\texttt{i}}(\texttt{s}) = 1$

(no code)

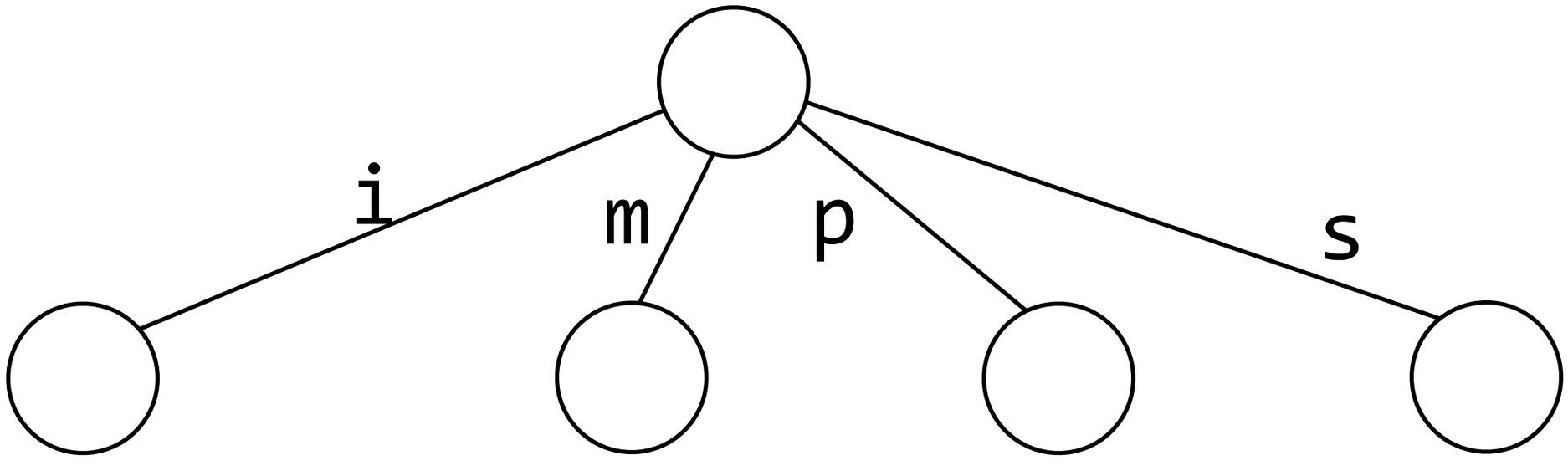
$C_{\texttt{p}}(\texttt{p}) = 0$

$C_{\texttt{p}}(\texttt{i}) = 1$

$C_{\texttt{s}}(\texttt{s}) = 0$

$C_{\texttt{s}}(\texttt{i}) = 1$

High-order entropy



$C_{\mathtt{i}}(\mathtt{p}) = 00$

$C_{\mathtt{i}}(\mathtt{m}) = 01$

$C_{\mathtt{i}}(\mathtt{s}) = 1$

(no code)

$C_{\mathtt{p}}(\mathtt{p}) = 0$ $C_{\mathtt{s}}(\mathtt{s}) = 0$

$C_{\mathtt{p}}(\mathtt{i}) = 1$ $C_{\mathtt{s}}(\mathtt{i}) = 1$

Can compress to
$\leq n(H_k(S) + 1)$ bits

Switching between many codes adds overhead

High-order entropy

H_k encoding reaches into the string, extracting "structure" needed to compress well

k balances compression with overhead

Grouping principle at play

H_0-based methods are simpler, faster, require less memory, but can't find as much structure as H_k

...or...can they? *Order* to the rescue