On the density of languages representing finite set partitions

Nelma Moreira Rogério Reis

Technical Report Series: DCC-04-07

Departamento de Ciência de Computadores – Faculdade de Ciências

&

Laboratório de Inteligência Artificial e Ciência de Computadores

Universidade do Porto

Rua do Campo Alegre, 823 4150 Porto, Portugal

Tel: +351+2+6078830 – Fax: +351+2+6003654

http://www.ncc.up.pt/fcup/DCC/Pubs/treports.html

On the density of languages representing finite set partitions

Nelma Moreira Rogério Reis
{nam,rvr}@ncc.up.pt
DCC-FC & LIACC, Universidade do Porto
R. do Campo Alegre 823, 4150 Porto, Portugal

August 2004

Abstract

We present a family of regular languages representing partitions of a set $\mathbb{N}_n = \{1, \ldots, n\}$ in less or equal c parts. We determine explicit formulas for the density of those languages and their relationship with well known integer sequences involving Stirling numbers of second kind. We also determine their limit frequency. This work was motivated by computational representations of the configurations of some numerical games.

1 The languages L_c

Consider a game where natural numbers are to be placed, by increasing order, in a fixed number of columns, subject to some specific constraints. In these games column's order is irrelevant. Game configurations can be seen as sequences of columns where the successive integers can be placed. For instance, the string

$$11213$$

stands for a configuration where 1, 2, 4 were placed in the first column, 3 was placed in the second and 5 was placed in the third. Because column order is irrelevant, and to have a unique representation for each configuration, it is illegal to place an integer in the k column if the $(k-1)$th is still empty, for any $k \geq 1$. Blanchard and al.[BHR04] and Reis and al.[RMP04], used this kind of representation for the possible configurations of sum-free games.

Given c columns, let $\mathbb{N}_c = \{1, \ldots, c\}$. We are interested in studying the set of game configurations as strings in $(\mathbb{N}_c)^\star$, i.e, in the set of finite sequences of elements of $\mathbb{N}_c$. We characterise game configurations by the following language $L_c \subset (\mathbb{N}_c)^\star$:

$$L_c = \{a_1 a_2 \cdots a_k \in (\mathbb{N}_c)^\star \mid \forall i \in \mathbb{N}_k, a_i \leq max\{a_1, \ldots, a_{i-1}\} + 1\}.$$

For $c = 4$, there are only 15 strings in L_c of length 4, instead of the total possible 256 in $(\mathbb{N}_4)^4$:

1111	1112	1121	1122	1123
1211	1212	1213	1221	1222
1223	1231	1232	1233	1234

We are going to show that L_c are regular languages. Given a finite set Σ, a regular expression (r.e.) α over Σ represents a (regular) language $L(\alpha) \subseteq \Sigma^\star$ and is inductively defined by: $\emptyset$ is a r.e and $L(\emptyset) = \emptyset$, ϵ (empty string) is a r.e and $L(\epsilon) = \{\epsilon\}$; $a \in \Sigma$ is a r.e and $L(a) = \{a\}$; if α_1 and α_2 are r.e., $\alpha_1 + \alpha_2$, $\alpha_1\alpha_2$ and $\alpha^\star$ are r.e., respectively with $L(\alpha_1 + \alpha_2) = L(\alpha_1) \cup L(\alpha_2)$, $L(\alpha_1\alpha_2) = L(\alpha_1)L(\alpha_2)$ and $L(\alpha_1{}^\star) = L(\alpha_1)^\star$ [HMU00]. A regular expression α is unambiguous if for each $w \in L(\alpha)$ there is only one path through α that matches w.

Theorem 1. *For all $c \geq 1$, L_c is a regular language.*

Proof. For $c = 1$, we have $L_1 = L(11^\star)$. We define by induction a family of regular expressions:

$$\alpha_1 = 11^\star, \tag{1}$$

$$\alpha_c = \alpha_{c-1} + \bigotimes_{j=1}^{c} j(1 + \cdots + j)^\star, \tag{2}$$

where $\otimes$ stand for the concatenation operator.

It is trivial to see that,

$$\alpha_c = \bigoplus_{i=1}^{c} \bigotimes_{j=1}^{i} j(1 + \cdots + j)^\star, \tag{3}$$

where $\oplus$ stands for the union operator. For instance, α_4 is

$$11^\star + 11^\star 2(1+2)^\star + 11^\star 2(1+2)^\star 3(1+2+3)^\star + 11^\star 2(1+2)^\star 3(1+2+3)^\star 4(1+2+3+4)^\star.$$

For any $c \geq 1$, we prove that

$$L_c = L(\alpha_c).$$

$L_c \supseteq L(\alpha_c)$**:** If $x \in L(\alpha_c)$ it is obvious that $x \in L_c$

$L_c \subseteq L(\alpha_c)$**:** By induction on the length of $x \in L_c$: If $|x| = 1$ than $x \in L(\alpha_1) \subseteq L(\alpha_c)$. Suppose that for any string x of length $\leq n$, $x \in L(\alpha_c)$. Let $|x| = n + 1$ and $x = ya$, where $a \in \mathbb{N}_c$ and $y \in L(\alpha_c)$. Let $c' = \max\{a_i \mid a_i \in y\}$. If $c' = c$, obviously $x \in L(\alpha_c)$. If $c' < c$, then $y \in L(\alpha_{c'})$, and $x \in L(\alpha_{c'+1}) \subseteq L(\alpha_c)$.

□

2 Counting the strings of L_c

The density of a language L over Σ, $\rho_L(n)$, is the number of strings of length n that are in L, i.e,

$$\rho_L(n) = |L \cap \Sigma^n|.$$

In particular, the density of L_c is

$$\rho_{L_c}(n) = |L_c \cap \mathbb{N}_c^n|.$$

We use generating functions to determine a closed form for $\rho_{L_c}(n)$. Recall that, a (ordinary) generating function for a sequence $\{a_n\}$ is a formal series [GKP94]

$$G(z) = \sum_{i=0}^{\infty} a_n z^n.$$

If $A(z)$ and $B(z)$ are generating functions for the density functions of the languages represented by unambiguous regular expressions A and B, and $A + B$, AB and $A^\star$ are also unambiguous r.e., we have that $A(z) + B(z)$, $A(z)B(z)$ and $\dfrac{1}{1 - A(z)}$, are the generating functions for the density functions of the corresponding languages (see [SF96], page 378).

As α_c are unambiguous regular expressions, from (3), we obtain the following generating function for $\{\rho_{L_c}(n)\}$:

$$T_c(z) = \sum_{i=1}^{c} \prod_{j=1}^{i} \frac{z}{(1-jz)} = \sum_{i=1}^{c} \frac{z^i}{\prod_{j=1}^{i}(1-jz)}.$$

Notice that

$$S_i(z) = \frac{z^i}{\prod_{j=1}^{i}(1-jz)}$$

are the generating functions for the Stirling numbers of second kind

$$S(n,i) = \frac{1}{i!}\sum_{j=0}^{i-1}(-1)^j \binom{i}{j}(i-j)^n,$$

i.e, the number of ways of partitioning a set of n elements into i nonempty sets [GKP94].

So, a closed form for the density of L_c, $\rho_{L_c}(n)$, is given by

$$\rho_{L_c}(n) = \sum_{i=1}^{c} S(n,i). \tag{4}$$

In Table 2 we present $\rho_{L_c}(n)$, for $c = 1..8$ and $n = 1..13$. For some sequences, we also indicate the correspondent number in Sloane's On Line Encyclopedia of Integer Sequences [Slo03]. The closed forms were calculated using the Maple computer algebra system [Hec03].

Moreover we can express (4) as a generic linear combination of n powers of i, for $i \in \mathbb{N}_c$. Let $S^j(n,i)$ denote the jth term in the summation of a Stirling number $S(n,i)$, i.e,

$$S^j(n,i) = \frac{1}{i!}(-1)^j \binom{i}{j}(i-j)^n.$$

Lemma 1. *For all $n, i \geq 0$,*

$$S^0(n,i) = -S^1(n,i+1). \tag{5}$$

Proof.

$$\begin{aligned} S^1(n,i+1) &= \frac{1}{(i+1)!}(-1)\binom{i+1}{1} i^n \\ &= (-1)\frac{1}{i!} i^n \\ &= -S^0(n,i). \end{aligned}$$

□

Applying (5) in the summation (4) of $\rho_{L_c}(n)$, each term $S(n,i)$ simplifies the subterm $S^1(n,i)$ with the subterm $S^0(n,i-1)$ of $S(n,i)$, for $i \geq 2$. So, we have

$$\begin{aligned} \rho_{L_1}(n) &= S^0(n,1), \\ \rho_{L_1}(n) &= S^0(n,2), \\ \rho_{L_c}(n) &= S^0(n,c) + \sum_{i=3}^{c}\sum_{j=2}^{i-1} S^j(n,i), && \text{for } c > 2 \\ &= \frac{c^n}{c!} + \sum_{i=3}^{c}\sum_{j=2}^{i-1} S^j(n,i), && \text{for } c > 2, \end{aligned}$$

c	$\rho_{L_c}(n)$	OISE
1	1 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...	
2	$\frac{1}{2}\,2^n$ 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, ...	
3	$\frac{1}{6}\,3^n + \frac{1}{2}$ 1, 2, 5, 14, 41, 122, 365, 1094, 3281, 9842, 29525, 88574, 265721, ...	A007051
4	$\frac{1}{24}\,4^n + \frac{1}{4}\,2^n + \frac{1}{3}$ 1, 2, 5, 15, 51, 187, 715, 2795, 11051, 43947, 175275, 700075, 2798251, ...	A007581
5	$\frac{1}{120}\,5^n + \frac{1}{12}\,3^n + \frac{1}{6}\,2^n + \frac{3}{8}$ 1, 2, 5, 15, 52, 202, 855, 3845, 18002, 86472, 422005, 2079475, 10306752, ...	A056272
6	$\frac{1}{720}\,6^n + \frac{1}{48}\,4^n + \frac{1}{18}\,3^n + \frac{3}{16}\,2^n + \frac{11}{30}$ 1, 2, 5, 15, 52, 203, 876, 4111, 20648, 109299, 601492, 3403127, 19628064, ...	A056273
7	$\frac{1}{5040}\,7^n + \frac{1}{240}\,5^n + \frac{1}{72}\,4^n + \frac{1}{16}\,3^n + \frac{11}{60}\,2^n + \frac{53}{144}$ 1, 2, 5, 15, 52, 203, 877, 4139, 21110, 115179, 665479, 4030523, 25343488 ...	
8	$\frac{1}{40320}\,8^n + \frac{1}{1440}\,6^n + \frac{1}{360}\,5^n + \frac{1}{64}\,4^n + \frac{11}{180}\,3^n + \frac{53}{288}\,2^n + \frac{103}{280}$ 1, 2, 5, 15, 52, 203, 877, 4140, 21146, 115929, 677359, 4189550, 27243100, ...	

Table 1: Density functions of L_c, for $c = 1..8$.

and, if the sums are rearranged such as that $i = k + j$, we have

$$\rho_{L_c}(n) = \frac{c^n}{c!} + \sum_{k=1}^{c-2}\sum_{j=2}^{c-k} S^j(n, k+j), \tag{6}$$

where

$$\begin{aligned} S^j(n, k+j) &= \frac{k^n}{(k+j)!}(-1)^j \binom{k+j}{j} && (7)\\ &= \frac{k^n}{k!j!}(-1)^j. && (8) \end{aligned}$$

Replacing (8) into (6) we get

$$\rho_{L_c}(n) = \frac{c^n}{c!} + \sum_{k=1}^{c-2}\frac{k^n}{k!}\left(\sum_{j=2}^{c-k}\frac{(-1)^j}{j!}\right). \tag{9}$$

In equation (9), the coefficients of k^n, $1 \le k \le c$, can be calculated using the following recurrence relation:

$$\begin{aligned} \gamma_1^1 &= 1;\\ \gamma_1^c &= \gamma_1^{c-1} + \frac{(-1)^{c-1}}{(c-1)!}, \text{ for } c > 1;\\ \gamma_k^c &= \frac{\gamma_{k-1}^{c-1}}{k}, \text{ for } c > 1 \text{ and } 2 \le k \le c. \end{aligned}$$

And, we obtain

Theorem 2. *For all $c \ge 1$,*

$$\rho_{L_c}(n) = \sum_{k=1}^{c}\gamma_k^c k^n. \tag{10}$$

Finally, the frequency of strings in L_c can be compared with that of $(\mathbb{N}_c)^\star$. Notice that

$$\rho_{(\mathbb{N}_c)^\star}(n) = c^n.$$

So, as $\lim_{n\to\infty}\left(\frac{k}{c}\right)^n = 0$, we have

Theorem 3. *For all $c \ge 1$,*

$$\lim_{n\to\infty}\frac{\rho_{L_c}(n)}{\rho_{(\mathbb{N}_c)^\star}(n)} = \frac{1}{c!}.$$

3 A bijection between strings of L_c and partitions of finite sets

The connection between the density of L_c and Stirling numbers of second kind is not accidental. Each string of L_c with length n corresponds to a partition of $\mathbb{N}_n$ with no more than c parts. This correspondence can be defined as follows.

Let $a_1a_2\cdots a_n$ be a string of L_c. This string corresponds to the partition $\{A_j\}_{j\in\mathbb{N}_{c'}}$ of $\mathbb{N}_n$ with $c' = \max\{a_1, \ldots, a_n\}$, such that for each $i \in \mathbb{N}_n$, $i \in A_j$.

For example, the string 1123 corresponds to the partition $\{\{1,2\},\{3\},\{4\}\}$ of $\mathbb{N}_4$ into 3 parts.

This defines a bijection. That each string corresponds to a unique partition is obvious. Given a partition $\{A_j\}_{j\in\mathbb{N}_{c'}}$ of $\mathbb{N}_n$ with $c' \le c$, we can construct the string $b_1\cdots b_n$, such that for $i \in \mathbb{N}_n$, $b_i = j$ if $i \in A_j$.

For the partition $\{\{1,2\},\{3\},\{4\}\}$ we obtain 1123.

4 Counting the strings of L_c of length equal or less than a certain value

Although strings of L_c of arbitrary length represent game configurations, for computational reasons[1] we consider all game configurations with the same length padding with zeros the positions of integers not yet in one of the c columns. In this way, we obtain the languages $L_c^0 = L_c\{0^\star\}$.

So, to determine the number of strings of length equal or less than n that are in L_c, is tantamount to determine the density of $L_c\{0^\star\}$, i.e

$$\rho_{L_c^0}(n) = |L_c^0 \cap (\{0\} \cup \mathbb{N}_c)^n|.$$

As seen in Section 2, and because $L_c\{0^\star\} = L(\alpha_c 0^\star)$, the generating function $T'_c(z)$ of $\rho_{L_c^0}(n)$, can be obtained using the generating function for $\rho_{L_c}(n)$, $T_c(z)$, and one for $\rho_{\{0\}^\star}(n)$. For the latter we have just

$$\frac{1}{1-z}.$$

So, the generating function for $\{\rho_{L_c^0}(n)\}$ is

$$T'_c(z) = T_c(z)\frac{1}{1-z} = \sum_{i=1}^{c} \frac{z^i}{(1-z)\prod_{j=1}^{i}(1-jz)}$$

and a closed form for $\rho_{L_c^0}(n)$ is (as excepted)

$$\rho_{L_c^0}(n) = \sum_{m=1}^{n}\sum_{i=1}^{c} S(m,i), \tag{11}$$

where m starts in 1 because $S(m,i) = 0$, for $i > m$.

Using expression (9) in (11) we have

$$\begin{aligned}
\rho_{L_1^0}(n) &= n,\\
\rho_{L_2\{0\}^\star}(n) &= 2^n - 1,\\
\rho_{L_c^0}(n) &= \sum_{m=1}^{n}\left(\frac{c^m}{c!} + \sum_{k=1}^{c-2}\frac{k^m}{k!}\sum_{j=2}^{c-k}\frac{(-1)^j}{j!}\right), \qquad \text{for } c > 2\\
&= \frac{c^{n+1}-c}{(c-1)c!} + n\sum_{j=2}^{c-1}\frac{(-1)^j}{j!} + \sum_{k=2}^{c-2}\frac{k^{n+1}-k}{(k-1)k!}\left(\sum_{j=2}^{c-k}\frac{(-1)^j}{j!}\right)\\
&= \frac{c^n-1}{(c-1)(c-1)!} + n\sum_{j=2}^{c-1}\frac{(-1)^j}{j!} + \sum_{k=2}^{c-2}\frac{k^n-1}{(k-1)(k-1)!}\left(\sum_{j=2}^{c-k}\frac{(-1)^j}{j!}\right).
\end{aligned}$$

If we use the equation (10) we obtain

Theorem 4. *For all $c \geq 1$,*

$$\rho_{L_c^0}(n) \quad = \quad n\gamma_1^c + \sum_{k=2}^{c}\frac{\gamma_k^c(k^{n+1}-k)}{k-1}.$$

[1] The data structures are arrays of fixed length.

Proof.

$$\begin{aligned}\rho_{L_c^0}(n) &= \sum_{m=1}^{n}\sum_{k=1}^{c}\gamma_k^c k^m \\ &= \sum_{k=1}^{c}\gamma_k^c\sum_{m=1}^{n}k^m \\ &= n\gamma_1^c+\sum_{k=2}^{c}\frac{\gamma_k^c(k^{n+1}-k)}{k-1}.\end{aligned}$$

□

In the Table 2 we present the values of $\rho_{L_c^0}(n)$, for $c=1..6$ and $n=1..13$.

c	$\rho_{L_c^0}(n)$	OISE
1	n 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, . . .	
2	2^n-1 1, 3, 7, 15, 31, 63, 127, 255, 511, 1023, 2047, 4095, 8191, ...	A000225
3	$\frac{1}{4}\,3^n+\frac{1}{2}\,n-\frac{1}{4}$ 1, 3, 8, 22, 63, 185, 550, 1644, 4925, 14767, 44292, 132866, 398587, . . .	A047926
4	$\frac{1}{18}\,4^n+\frac{1}{2}\,2^n+\frac{1}{3}\,n-\frac{5}{9}$ 1, 3, 8, 23, 74, 261, 976, 3771, 14822, 58769, 234044, 934119, 3732370, . . .	
5	$\frac{1}{96}\,5^n+\frac{1}{8}\,3^n+\frac{1}{3}\,2^n+\frac{3}{8}\,n-\frac{15}{32}$ 1, 3, 8, 23, 75, 277, 1132, 4977, 22979, 109451, 531456, 2610931, 12917683, . . .	
6	$\frac{1}{600}\,6^n+\frac{1}{36}\,4^n+\frac{1}{12}\,3^n+\frac{3}{8}\,2^n+\frac{11}{30}\,n-\frac{439}{900}$ 1, 3, 8, 23, 75, 278, 1154, 5265, 25913, 135212, 736704, 4139831, 23767895, . . .	
7	$\frac{1}{4320}\,7^n+\frac{1}{192}\,5^n+\frac{1}{54}\,4^n+\frac{3}{32}\,3^n+\frac{11}{30}\,2^n+\frac{53}{144}\,n-\frac{31}{64}$ 1, 3, 8, 23, 75, 278, 1154, 5265, 25913, 135212, 736704, 4139831, 23767895, . . .	
8	$\frac{1}{35280}\,8^n+\frac{1}{1200}\,6^n+\frac{1}{288}\,5^n++\frac{1}{48}\,4^n+\frac{11}{120}\,3^n+\frac{53}{144}\,2^n+\frac{103}{280}\,n-\frac{57023}{117600}$ 1, 3, 8, 23, 75, 278, 1155, 5295, 26441, 142370, 819729, 5009279, 32252379, . . .	

Table 2: Density functions of L_c^0, for $c=1..8$.

Finally, we determine the frequency of strings in L_c^0 in $(\mathbb{N}_c)^\star\{0\}^\star$, for $c>1$. Notice that

$$\rho_{(\mathbb{N}_c)^\star\{0\}^\star}(n)=\frac{c^{n+1}-1}{c-1}.$$

as it is a sum of the first n terms of a geometric progression of ratio c.

We have,

Theorem 5. *For all $c \geq 1$,*

$$\lim_{n\to\infty} \frac{\rho_{L_c^0}(n)}{\rho_{(\mathbb{N}_c)^\star\{0\}^\star}(n)} = \frac{1}{c!}.$$

Proof.

$$\begin{aligned}
\lim_{n\to\infty} \frac{\rho_{L_c^0}(n)}{\rho_{(\mathbb{N}_c)^\star\{0\}^\star}(n)} &= \lim_{n\to\infty}\left(\frac{(c^{n+1}-c)(c-1)}{(c-1)c!(c^{n+1}-1)} + \frac{n(c-1)}{c^{n+1}-1}\sum_{j=2}^{c-1}\frac{(-1)^j}{j!}\right) \\
&\quad + \lim_{n\to\infty}\left(\sum_{k=2}^{c-2}\frac{(k^{n+1}-k)(c-1)}{(k-1)k!(c^{n+1}-1)}\left(\sum_{j=2}^{c-k}\frac{(-1)^j}{j!}\right)\right) \\
&= \frac{1}{c!}\left(\lim_{n\to\infty}\frac{1}{1-\frac{1}{c^{n+1}}} - \lim_{n\to\infty}\frac{c}{c^{n+1}-1}\right) \\
&\quad + \sum_{j=2}^{c-1}\frac{(-1)^j}{j!}\lim_{n\to\infty}\left(\frac{n(c-1)}{c^{n+1}-1}\right) \\
&\quad + \sum_{k=2}^{c-2}\frac{(c-1)}{(k-1)(k-1)!}\sum_{j=2}^{c-k}\frac{(-1)^j}{j!}\lim_{n\to\infty}\left(\frac{(\frac{k}{c})^n}{c-\frac{1}{c^{n+1}}} - \frac{1}{c^{n+1}-1}\right) \\
&= \frac{1}{c!}.
\end{aligned}$$

□

5 Conclusion

In this note we presented a family of regular languages representing finite set partitions and studied their densities. Although it is well known that the number of partitions of a set of n elements into (no more than) c nonempty sets is given by (the sum of) Stirling numbers of second kind, $S(n, c)$, we determined explicit formulas for the number of partitions of a set of n elements with no more than c. We also determined the limit density of those languages.

Acknowledgement

We thank Miguel Filgueiras for helpful comments.

References

[BHR04] Peter Blanchard, Frank Harary, and Rogério Reis. Partitions into sum-free sets. *To appear*, 2004.

[GKP94] Roland Graham, Donald Knuth, and Oren Patashnik. *Concrete Mathematics*. AW, 1994.

[Hec03] André Heck. *Introduction to Maple*. SV, 3rd edition, 2003. http://remote.science.uva.nl/

[HMU00] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison Wesley, 2nd edition, 2000.

[RMP04] Rogério Reis, Nelma Moreira, and João Pedro Pedroso. Educated brute-force to get h(4). Technical Report DCC-2004-04, DCC-FC& LIACC, Universidade do Porto, July 2004.

[SF96] Robert Sedgewick and Philippe Flajolet. *Analysis of Algorithms*. AW, 1996.

[Slo03] N.J.A. Sloane. The On-line Encyclopedia of Integer Sequences, 2003. http://www.research.att.com/∼njas/sequences.

2000 *Mathematics Subject Classification*: Primary 05A15; Secondary 05A18, 11B73, 68Q45.

Keywords: Partions of sets, Stirling numbers, regular languages, enumeration

---

(Concerned with sequences A00705 A007581 A056272 A056273 A000225 A047926.)