Three remarks on Matula numbers

Albert Burgos
Universidad de Murcia
30100 Espinardo, Murcia, Spain
e-mail: albert@um.es

Abstract

In SIAM Review 10, page 273, D. W. Matula described a bijection between \( \mathbb{N} \) and the set of topological rooted trees; the number is called the Matula number of the rooted tree. The Gutman-Ivić-Matula (GIM) function \( g(n) \) computes the number of edges of the unique tree with Matula number \( n \). Since there is a prefix-free code for the set of prime numbers such that the codelength of each prime \( p \) is \( 2g(p) \), we show how some properties of the GIM function can be obtained trivially from coding theorems.

Key words: Matula numbers, Kraft’s inequality, Shannon’s entropy.

Introduction

\( \mathbb{N} \) stands for the set of natural numbers, \( \mathbb{W} \) for the set of whole numbers \( \mathbb{N} \cup \{0\} \), \( \mathcal{T} \) for the set of finite and undirected rooted trees, and \( \mathbb{T} \) for the set of topological rooted trees (i.e. all equivalence classes of undirected rooted trees where the equivalence is the natural isomorphism). Let \( S \subset \mathbb{N} \) (respectively \( \mathbb{W} \)) be a set defined by property such that every natural number (respectively whole number) has a unique decomposition as a multiset of \( S \). Since the set of integers is totally ordered, then \( S \) is totally ordered. If the least element in \( S \) is greater than 1 (resp. 0), the index of an element \( s_i \in S \) is less than \( s_i \), i.e. \( s_i \in S \) implies \( i < s_i \). This leads to a recursive map from the naturals into \( \mathbb{T} \). We can easily construct examples of such property by setting \( S = \{b^n|n \geq 0\} \) (for some base \( b > 1 \)) or \( S = \{n!|n \geq 1\} \). In these two cases, decomposition builds on the addition operation on \( \mathbb{W} \). However, the prototype of all \( S \) defined as above

---

1 Preliminary draft, please do not quote without the author permission. I thank Maríα F. Morales for helpful comments. I also thank Regis de la Bretèche who have pointed out to me an important reference, from which I borrow the term of GIM function. Of course, all errors and/or omissions are my own. Financial support from the Spanish Ministry of Education through the project ECO2010-21624 is gratefully acknowledged.
is of course the set of all primes \( \mathbb{P} = \{ p(n) | n \geq 1 \} \), for which decomposition is the product of elements in \( \mathbb{P} \). In this case, the fundamental theorem of arithmetic tells us that the recursive map is a denumeration of \( \mathbb{T} \), as was noted independently by Matula [1] and Göbel [2], which give an explicit construction of the map and its inverse. Thus, any statistics on rooted trees may be assigned to a natural number. This has been done for several statistics starting with the pioneering work of Gutman, Ivić and Elk [3], with follow-ups by Gutman and Yeh [4], Gutman and Ivić [5], [6], and Deutsch [7], among others. In particular, Gutman, Ivić and Elk ask themselves how many edges does a natural number have, proving that the solution is given by the unique function completely additive \( g \), such that, for any \( n \in \mathbb{N} \),

\[
g(p(n)) = 1 + g(n).
\]  

This function, dubbed by La Bretèche and Tenenbaum in [8] as the Gutman-Ivić-Matula (GIM) function, has a simple translation in terms of coding theory: It represents the semi-length of the codewords when natural numbers are encoded a Dyck alphabet via the Matula bijection. By exploiting this relation we can apply basic results on information entropy to provide non-asymptotic results on its behavior.

1 Preliminaries

Write \( \pi(n) \) for the number of primes less than or equal to \( n \), and \( p(n) \) for the \( n \)-th prime in ascending order. Given \( n \in \mathbb{N}_{\geq 2} \), if the prime factorization of \( n \) is \( f_1 \cdot f_2 \cdots \cdot f_m \) and \( \{ \circ \} \) denotes the rooted tree consisting of a single node, then the function \( \tau : \mathbb{N} \mapsto \mathbb{T} \) is defined as a recursion:

1. \( \tau(1) = \{ \circ \} \).
2. For \( n \geq 2 \), \( \tau(n) \) is the tree in which the root is adjacent to the roots of \( \tau(\pi(f_i)) \) for \( 1 \leq i \leq m \).

The map \( \tau \) is a bijection. Its inverse is defined as follows:

1. \( \tau^{-1}(\circ) = 1 \).
2. If the root of tree \( t \) is adjacent to subtrees \( t_1, t_2, \ldots, t_m \), then

\[
\tau^{-1}(t) = \prod_{i=1}^{m} p(\tau^{-1}(t_i)).
\]

A prime factorization of a number is in canonical order when the primes are presented in nondecreasing order. An analogue for rooted trees goes as follows: If \( \tau(n) = t \), the rooted tree \( t \) is presented canonically when:
(1) The rooted trees $t_1, t_2, \ldots, t_m$, corresponding to the factors $f_1, f_2, \ldots, f_m$, respectively, are presented from left to right.

(2) Each rooted tree $t_i$ is presented canonically.

Notice that $n \in \mathbb{P}$ iff $\tau(n) \in T_p$, where $T_p \subset T$ is the set of all planted rooted trees.\footnote{i.e. rooted trees for which the root has degree 1.}

Let $(\mathbb{N}, \cdot, 1)$ denote the commutative monoid of the positive integers under product. Let the merging of rooted trees $t_1$ and $t_2$, denoted, $t_1 \land t_2$, to be the labeled tree that results from identifying their roots. Now let $(T, \land, \circ)$ be the commutative monoid of rooted trees under product, and let $p(t)$ be the rooted tree whose root is adjacent to subtree $t$. Since $\tau$ is a bijection, there is a isomorphism between $(\mathbb{N}, \cdot, 1)$ and $(T, \land, \circ)$, which trivially extends to an isomorphism of the algebra $\mathcal{N} = \langle \mathbb{N}, \cdot, p, 1 \rangle$ onto the algebra $T = \langle T, \land, p, \circ \rangle$.

2 The Matula code

In what follows, we are motivated to use the algebra $T$ to encode the positive integers. Any tree can be written as a string of symbols from the set $\{p, \circ, \land, (, )\}$. In any string representing a tree presented in canonical form, $p$ is always followed by $($, the operator $\land$ is always followed by $p$, and $\circ$ appears only between $($ and $)$. Hence, no information is lost if we drop all the $p$'s, $\circ$'s, and $\land$'s. For example, the rooted tree (b) in the figure above is represented as $p(p(p(\circ) \land p(\circ))))$, which can be simply written as $((())())$.

We need now to introduce some relevant formalism. I shall call alphabet a given set $\Sigma$ such that $2 \leq |\Sigma| < +\infty$. Elements of $\Sigma$ are called symbols. A string or word over $\Sigma$ is any finite sequence of symbols from $\Sigma$. The length of a string is the number of symbols in the string (the length of the sequence) and can be any non-negative integer. The empty string is the unique string over $\Sigma$ of length 0, and is denoted $e$. The set of all strings over $\Sigma$ of length $t$ is denoted $\Sigma^t$. (Note that $\Sigma^0 = \{e\}$ for any $\Sigma$.) The set of all finite-length strings over $\Sigma$ is the Kleene closure of $\Sigma$, denoted $\Sigma^*$. For any two strings $s$ and $s'$ in $\Sigma^*$, their concatenation is defined as the sequence of symbols in $s$ followed by the sequence of symbols in $s'$, and is denoted $ss'$. The empty string serves as the identity element; for any string $s$, $es = se = s$. Therefore, the set $\Sigma^*$ and the concatenation operation form a monoid, the free monoid generated by $\Sigma$. In addition, the length function defines a monoid homomorphism $\ell : \Sigma^* \to \mathbb{N}$. Given a set $S$ with Kleene closure $S^*$, a code is a function $c : S \to \Sigma^*$. The elements of $c(S)$ will be referred to as the codewords, and $c$ is said to be a $|\Sigma|$-ary code. In this paper we shall assume that any code $c$ satisfies that if
s, s′, ss′ ∈ c(S) and s ≠ e, then s′ = e (i.e. the code is prefix-free).

In accordance with the above definitions, by setting Σ = { (, ) } and S = P, it is clear that τ induces a code, which we dub the Matula code, denoted c_M.

3 Concluding remarks

The following are results of the direct translation of the coding approach into properties of the Gutman-Ivić-Matula arithmetic function.

In Gutman, Ivić and Elk [3], Gutman and Yeh [4], Gutman and Ivić [5], and Gutman and Ivić [6] is introduced an arithmetic function capturing some graph theoretic properties of τ:

Definition 1 Let g : N → C be the unique function completely additive such that, for any n ∈ N,

\[ g(p(n)) = 1 + g(n). \]  \hspace{1cm} (2)

Then, we have:

Lemma 1 For all p ∈ P, \( \ell \circ c_M(p) = 2g(p) \).

Proof. It trivially follows from Theorem 3(a) in Gutman and Ivić [5]. ■

Then, we have,

Conclusion 1: \( M_P = \sum_{p \in P} 1/4^g(p) < 1/2 \).

Proof. By Claim 1 and Kraft’s inequality. ■

And,

Conclusion 2: \( M_N = \sum_{n \in N} 1/4^g(n) = M_P/4 < 2 \).

Proof. By Conclusion 1 and equation (2). ■

Thus, \( M_P \) is a probability, whereas \( M_P \) is not, because \( g(2) = 1 \) and therefore \( M_P > 1/4 \).

Bounds on g appear in Gutman and Ivić [5], and [6], which show, for all \( n \geq 7 \),

\[ \underline{g}(n) = \frac{\ln n}{\ln(\ln n)} \leq g(n) \leq \frac{3\ln n}{\ln 5} = \bar{g}(n). \]  \hspace{1cm} (3)

The bounds \( g \) and \( \bar{g} \) cannot be improved upon—i.e. there exist infinite values of \( n \) for which they are reached asymptotically. However, Conclusion 2 implies
that most values of $g$ are close to those of $\bar{g}$. Indeed, in [8], La Bretteche and Tenenbaum develop a method which evaluates the moments of $g$, showing that if $G(n) = \sum_{1 \leq i \leq n} g(i)$, then

$$G(n) = \phi n \ln n + O(n \ln(\ln n))$$

for some constant $\phi > 0$ (see [8] Theorem 1 and Theorem 3). Now, we shall see that Shannon’s source coding theorem [9] offers a proof of the fact that for all $n$, $G(n) > \phi n \ln n$ whenever $\phi \leq 1/\ln 4$.

Let $S$ be a finite subset of $\mathbb{N}$, $2^S$ the associated $\sigma$-algebra, and $\mu = (\mu(s))_{s \in S}$ a probability distribution on the elements of $S$. We shall denote by $S$ the discrete random variable taking values in $S$ according to $\mu$. Thus, the expectation for the length of $c_M(S)$ is

$$E[\ell \circ c_M(S)] = 2 \sum_{i=1}^{n} \mu(i) g(i). \quad (4)$$

Shannon’s coding theorem places a lower bound on this expected length. Namely, the ratio between the entropy of $S$, $H(S) = -\sum_{s \in S} \mu(s) \log \mu(s)$, and the cardinal of the target alphabet $\Sigma$. Therefore

$$E[\ell \circ c_{MG}(S)] \geq H(S) = -\sum_{s \in S} \mu(s) \log \mu(s), \quad (5)$$

i.e.

$$2 \sum_{s \in S} \mu(s) g(s) \geq \sum_{s \in S} \mu(s) \log \mu(s) \quad (6)$$

Trivially, $H(S) \leq \log |S|$ with equality if and only if $\mu$ is the uniform distribution. Thus, we get:

**Conclusion 3:** For all $n \in \mathbb{N}$,

$$G(n) \geq \frac{n}{\ln 4} \ln n. \quad (7)$$

**Proof.** Set $n \in \mathbb{N}$. If $S = [n]$ and $\mu(i) = 1/n$ for all $i \in S$, (3.6) yieldst $G(n) \geq n \log n/2 = n \ln n / \ln 4$. □

**References**

[1] D.W. Matula, A Natural Root Tree Enumeration by Prime Factorization, *SIAM Review*, 10: 273, 1968.

---

Here and thereafter we denote by log and ln the logarithms in bases 2 and $e$, respectively.
[2] F. Göbel, On a 1-1 correspondence between rooted trees and natural numbers, *Journal of Combinatorial Theory, Series B* **29**: 141-143, 1980.

[3] I. Gutman, A. Ivić, and S.B. Elk, Matula numbers for coding chemical structures and some of their properties, *Journal of the Serbian Chemical Society* **58** (3-4): 193-201, 1993.

[4] I. Gutman and Y.-N. Yeh, Deducing properties of trees from their Matula numbers, *Publications de l’Institut Mathématique, Nouvelle série* **53** (67) 17-22, 1982.

[5] I. Gutman and A. Ivić, Graphs with maximal and minimal Matula numbers, *Bulletin Académie Serbe des Sciences et des Arts (Classe des Sciences Mathématiques et Naturelles)* **107**: 65-74, 1994.

[6] ——, On Matula numbers, *Discrete Mathematics* **41**: 199-214, 1982.

[7] E. Deutsch, Rooted tree statistics from Matula numbers, *Discrete Applied Mathematics* **160** (15): 2314-2322, 2012.

[8] R. de La Bretèche and G. Tenenbaum, Sur certaines équations fonctionnelles arithmétiques, *Annales de l’institut Fourier* **50** (5): 1445-1505, 2000.

[9] C.E. Shannon, A Mathematical Theory of Communication, *Bell System Technical Journal*, **27** 379-423, 623-656, 1948.