Multiset and Set Decipherable Codes

F. Blanchet-Sadri and C. Morgan

F. Blanchet-Sadri and C. Morgan, "Multiset and Set Decipherable Codes." Computers and Mathematics with Applications: An International Journal, Vol. 41, No. 10/11, 2001, pp 1257-1262. doi:10.1016/S0898-1221(01)00096-7

Made available courtesy of Elsevier: http://www.elsevier.com/locate/camwa

Abstract:
We extend some results of Lempel and Restivo on multiset decipherable codes to set decipherable codes.

Keywords: Codes, Unique decipherability, Multiset decipherability, Dominoes, Set decipherability.

Article:

1. INTRODUCTION

In unique decipherable (UD) codes, different sequences of code words carry different information, In [1], Lempel introduces the notion of a multiset decipherable (MSD) code to handle some special problems in the transmission of information. Here the information of interest is the multiset of code words used in the encoding process so that order in which transmitted words are received is immaterial. In [2], Guzmán develops the concept of a set decipherable (SD) code, There it is the set of code words that is relevant information so the order and the multiplicity of words are immaterial.

The UD, MSD, and SD concepts coincide for two-word codes [1,3]. Lempel [1] conjectured that the UD and MSD concepts coincide for three-word codes or every MSD code of three words is a UD code, and Guzmán [3] conjectured that the UD, MSD, and SD concepts coincide for three- word codes, References [4-6] positively support these two conjectures. Lempel [1] constructs for $n \geq 4$, an n-word MSD code that is not UD or a proper MSD code.

The McMillan Sum for a code C over an alphabet A is given by

$$\mathrm{MS}(C) = \sum_{w \in C} |A|^{-|w|},$$

where A is the cardinality of the alphabet A and $|w|$ denotes the length of w. Every UD code C satisfies $\mathrm{MS}(C) \leq 1$ [7]. This inequality is known as Kraft's inequality and, intuitively, indicates that the words of a UD code cannot become "too short". In [1], Lempel conjectured that every MSD code satisfies Kraft's inequality. However, Restivo [8] showed that there exists an MSD code C such that $\mathrm{MS}(C) > 1$, and consequently, there exists an SD code C such that $\mathrm{MS}(C) > 1$. The resulting shorter average word-length of MSD codes is then a welcome trade-off for the weaker decipherability condition. This leaves open the possibility that there may exist situations in which MSD codes can provide greater efficiency in terms of word-lengths than UD codes.

In this paper, an n-word SD code that is not MSD or a proper SD code is constructed for $n \geq 4$. A result of Restivo [8], originally conjectured by Lempel [1], stating that no MSD code contains a full UD code as a proper subcode is extended to SD codes. Here a UD code C is called *full* if $\mathrm{MS}(C) = 1$.

2. UD, MSD, AND SD CODES

We now define precisely the three concepts of unique, multiset, and set decipherable codes.

Let A be a finite set that we call an *alphabet.* Its elements are called *letters. A word* over the alphabet A is a finite sequence of elements of *A*. The set of all words over *A* is denoted by A^*. The empty sequence, called the *empty word,* is denoted by *e.* The set of all nonempty words over *A* is denoted by A^+. *A code C* over *A* is a nonempty finite subset of A^+. The words in *C* are called *code words.* A *message* over *C* is a word in A^* that is a concatenation of code words. The sequence of these code words is a *decoding* or *factorization* of the message. The code *C* is called

- *uniquely decipherable* or UD, if every message over *C* has a unique factorization into code words,
- *multiset decipherable* or MSD, if any two factorizations of the same message over *C* yield the same multiset of code words,
- *set decipherable* or SD, if any two factorizations of the same message over *C* yield the same set of code words.

Let UD (respectively, MSD, SD) denote the class of all UD (respectively, MSD, SD) codes. It is clear that UD $\subseteq$ MSD $\subseteq$ SD and it has been shown that the two inclusions are strict. The code C_1 = {110, 101, 11011, 01110101} shows that the first inclusion is strict. In fact, the message

$$(110)(11011)(101)(01110101) = (11011)(01110101)(110)(101)$$

has two distinct factorizations into code words [1]. The code C_2 = {01, 10, 0010100,1001001} shows the strictness of the second inclusion. The message

$$(10)(01)(0010100)(10)(1001001) = (1001001)(01)(0010100)(10)(01)$$

has two distinct factorizations with distinct multisets of code words. This latter code is an instance of a complete list of proper MSD and proper SD four-word codes over {0,1} with code words of length less than or equal to 7 given by Guzmán [3]. It is decidable whether or not a code *C* is UD or MSD [9-12], respectively).

First, we give in Section 3 a brief overview of Head and Weber's domino technique [10], and then give in Section 4, an application of it by constructing proper SD codes.

3. A DOMINO TECHNIQUE

Let A be a finite alphabet and *C* a code over A. Guzmán suggested looking at the *simplified domino graph* and the *domino function* of *C.* The simplified domino graph of *C* is a subgraph of the domino graph of *C* defined in [10].

Let Prefix(*C*) be the set of all prefixes of words in *C,* and let *G* = (*V, E*) be the directed graph with vertex set

$$V = \left\{\mathbf{open}, \mathbf{close}, \binom{u}{\epsilon}\binom{\epsilon}{u} \middle| u \in Prefix(C)\{\epsilon\}\right\}$$

and with edge set E = $E_1 \cup E_2 \cup E_3 \cup E_4$, where

$$E_1 = \left\{\left(\mathbf{open}, \binom{\epsilon}{u}\right) \middle| u \in C\right\},$$
$$E_1 = \left\{\left(\binom{u}{\epsilon}, \mathbf{close}\right) \middle| u \in C\right\},$$

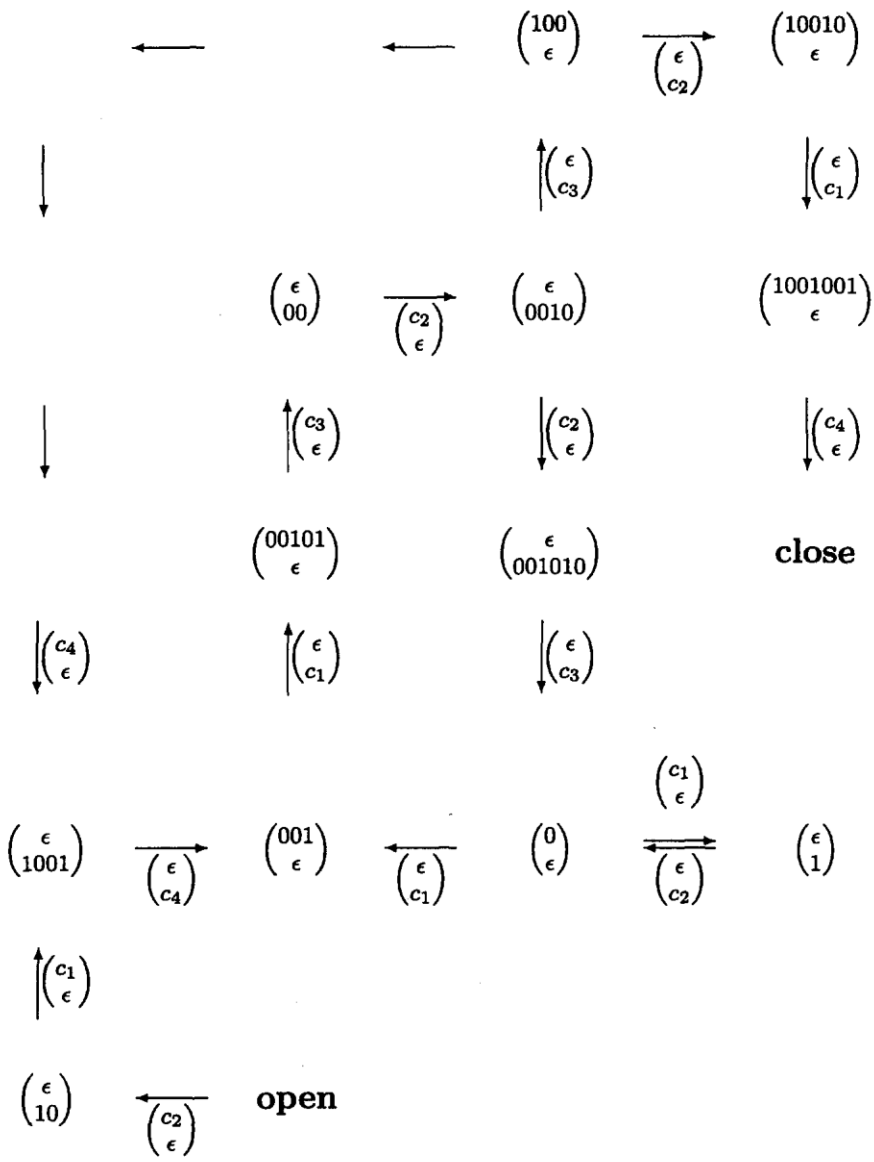

Figure 1. Simplified domino graph of $C_2 = \{01, 10, 0010100, 1001001\}$.

$$E_3 = \left\{ \left(\binom{u}{c}, \binom{u}{v}\epsilon \right), \left(\binom{\epsilon}{u}, \binom{\epsilon}{uv} \right) \middle| v \in C \right\},$$
$$E_4 = \left\{ \left(\binom{u}{\epsilon}, \binom{\epsilon}{v} \right), \left(\binom{\epsilon}{u}, \binom{v}{\epsilon} \right) \middle| uv \in C \right\}.$$

The *simplified domino graph* associated with C is the directed graph $G' = (V', E')$, where V' consists of **open, close,** and those vertices $v \in V$ such that there exists a path from **open** to **close** that goes through v, and E' consists of those edges $e \in E$ such that there exists a path from **open** to **close** going through e. The simplified domino graph of C is denoted by $G(C)$. The *domino function* associated with C is the mapping d from E to $\left\{ \binom{u}{\epsilon}, \binom{\epsilon}{u} | u \in C \right\}$ defined on

$$E_1 \; by \left(\textbf{open}, \binom{\epsilon}{u} \right) \mapsto \binom{u}{\epsilon},$$
$$E_2 \; by \left(\binom{u}{\epsilon}, \textbf{close} \right) \mapsto \binom{u}{\epsilon},$$

$$E_3\ by\ \left(\binom{u}{\epsilon},\binom{uv}{\epsilon}\right)\mapsto\binom{\epsilon}{v}\ and\ \left(\binom{\epsilon}{u},\binom{\epsilon}{uv}\right)\mapsto\binom{v}{\epsilon},$$
$$E_4\ by\ \left(\binom{u}{\epsilon},\binom{\epsilon}{u}\right)\mapsto\binom{uv}{\epsilon}\ and\ \left(\binom{\epsilon}{u},\binom{v}{\epsilon}\right)\mapsto\binom{\epsilon}{uv}.$$

The *domino* associated with an edge e of E is the domino $d(e)=\binom{d_1(e)}{d_2(e)}$. The function d induces mappings d_1 and d_2 from E to $C \cup \{\varepsilon\}$ also called domino functions. If $p = e_1 \ldots e_m$ is a path in G, the word $d(e_1)\ldots d(e_m)$ (respectively, $d_1(e_1)\ldots d_1(e_m)$, $d_2(e_1)\ldots d_2(e_m)$) will be denoted by $d(p)$ (respectively, $d_1(p)$, $d_2(p)$).

A path p in G from open to some vertex $\binom{u}{\epsilon}$ (respectively, $\binom{\epsilon}{u}$) is trying to find two factorizations of the same message over C into code words beginning with distinct code words. The decodings obtained so far are $d_1(p)$ and $d_2(p)$, The word $u \in A^*$ denotes the backlog of the first (respectively, second) decoding as against the second (respectively, first) one.

The following lemma states that the UD, MSD, and SD properties of a code C can be characterized in terms of its simplified domino graph $G(C)$ and the functions d_1 and d_2.

LEMMA 1.

- *$C \in$ UD if and only if* no *path exists in $G(C)$ from* **open** to **close** *[11].*
- *$C \in$ MSD if and only if all paths p in $G(C)$ from* **open** *to* **close** are *such* that $d_1(p)$ and $d_2(p)$ *have the same multiset of code words [10].*
- *$C \in$ SD if and only if all paths p in $G(C)$ from* **open** to **close** are *such that $d_1(p)$ and $d_2(p)$* have *the same set of code words [3].*

As an example, let us consider the code $C_2 = \{c_1, c_2, c_3, c_4\}$, where $c_1 = 01$, $c_2 = 10$, $c_3 = 0010100$, and $c_4 = 1001001$. Figure 1 gives the simplified domino graph of C_2 where each edge e is labelled by $d(e)$. Figure 2 gives the simplified domino graph of C_2 where each edge is relabelled by a number. This relabelling is useful in the sequel.

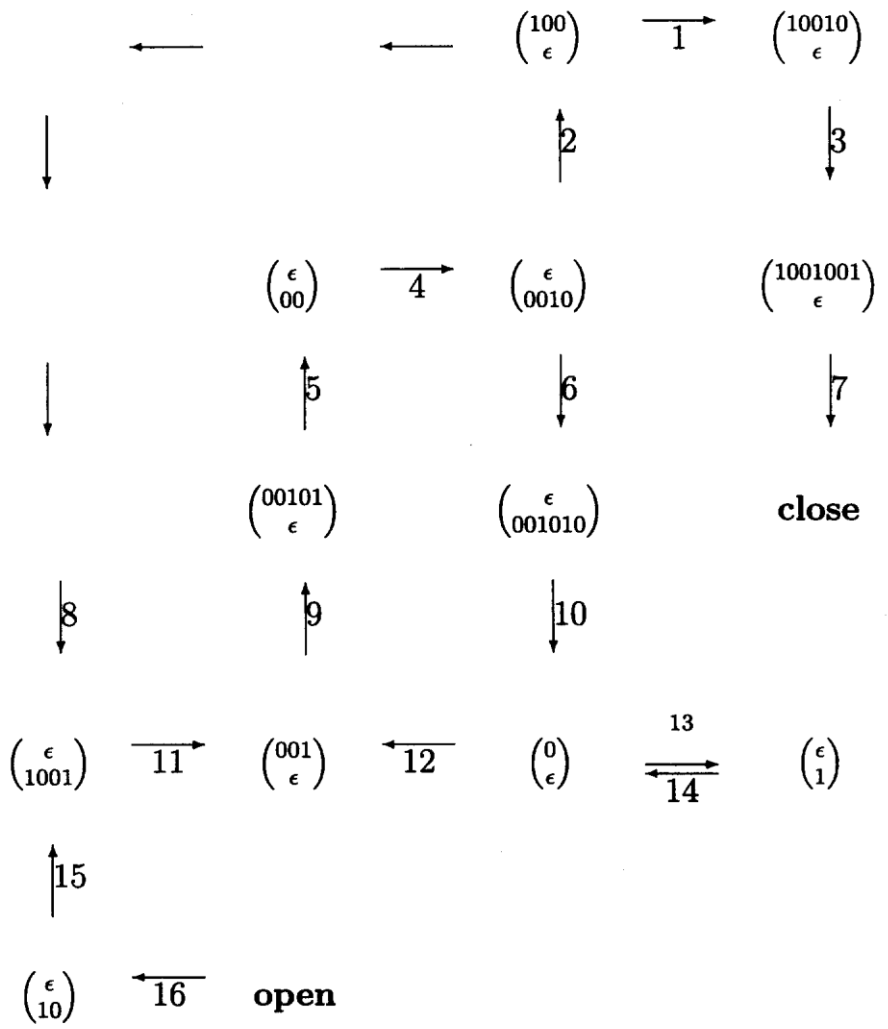

Figure 2.

In the next section, we use the fact that C_2 is a proper four-word SD code as a basis for building proper n-word SD codes for $n > 4$.

4. SD VERSIONS OF SOME RESULTS OF LEMPEL AND RESTIVO

We first show the existence of proper *n*-word SD codes for $n \geq 4$.

THEOREM 1, *Let* $\{d_1,\ldots,d_k\}$ *be a k-word prefix code, The code*

$$D_k = \{01, 10, 0010100, 1001001, 000001d_1, \ldots, 000001d_k\}$$

is a proper SD code of k + 4 words,

PROOF. First, we show that $C_2 = \{01, 10, 0010100, 1001001\} \in SD\backslash MSD$. The path

$$p = \textbf{open}, \binom{\epsilon}{10}, \binom{\epsilon}{1001}, \binom{001}{\epsilon}, \binom{00101}{\epsilon}, \binom{\epsilon}{00}, \binom{\epsilon}{0010}, \binom{100}{\epsilon}, \binom{10010}{\epsilon}, \binom{1001001}{\epsilon}, \textbf{close}$$

of Figure 1 (or the path 16, 15, 11, 9, 5, 4, 2, 1, 3, 7 of Figure 2) is from **open** to **close**. We see that $d_1(p) = c_2c_1c_3c_2c_4$ and $d_2(p) = c_4c_1c_3c_2c_1$ and so the message

$$10010010100101001001$$

has two distinct factorizations with distinct multisets of code words showing that $C_2 \notin$ MSD. To show that $C_2 \in$ SD, note that any path p from **open** to **close** contains at least the edges 16, 15, 11, 9, 5, 4, 2, 1, 3, and 7, and so $d_1(p)$ and $d_2(p)$ have the same set of code words $\{c_1, c_2, c_3, c_4\}$.

In order to prove the result, it suffices to show that $G(D_k) = G(C_2)$ and then $D_k \in SD\backslash MSD$. Referring to Figure 1, when trying to build $G(D_k)$, note that there is no edge from open to any $\binom{\epsilon}{000001d_i}$ since $\{d_1,\ldots,d_k\}$ is a prefix code, It is a simple matter to check that in $G(D_k)$, no edges other than the ones in $G(C_2)$ will be leaving $\binom{0}{\epsilon}$ or $\binom{\epsilon}{00}$. ■

We end with results on the McMillan Sum of SD codes.

THEOREM 2. *No SD code contains a full UD code as a proper subcode.*

PROOF. The proof is along the lines of the proof of the MSD version of this result given in [8]. Assume on the contrary that C is an SD code over an alphabet A containing a full UD code D as a proper subcode, and let $x \in C \setminus D$. By a known fact about UD codes [13], D is complete, and therefore, A^* is the set of factors of words in D^*. Since D is finite, D^* is regular and is accepted by a deterministic finite automaton $M = (Q, A, \delta, q_0, F)$. If $S \subseteq Q$ and $w \in A^*$, then Sw will denote the set $\{qw \mid q \in S\}$, where qw represents the state reached from q after reading w and often denoted by $\delta(q, w)$. Let n be the positive integer $\min_{w\in A^*} |Qw|$ and let u be such that $|Qu| = n$. Since D is complete, u is a factor of a word in D^* or there exist $v_1, v_2 \in A^*$ such that $v_1uv_2 = y \in D^*$, and consequently, $\delta(q_0, y) \in F$. Since $Qv_1u \subseteq Qu$, we have $|Qy| \leq |Qu|$, and therefore, $|Qy| = n$. Put $Q' = Qy$. Since $Q'y = Qyy \subseteq Qy = Q'$, it follows from the minimality of n that $Q'y = Q'$, and thus, y defines a permutation of Q'. There exists a positive integer ℓ such that y^ℓ is the identity permutation of Q' or $q'y^\ell = q'$ for all $q' \in Q'$. If $z = y^\ell xy^\ell$, then $Qz \subseteq Qyt$ and $Qz = Q' = Q'z$. Thus, for some positive integer m, we have $q'z^m = q'$ for all $q' \in Q'$. We prove that $z^m = (y^\ell xy^\ell)^m \in D^*$ by showing that $qz^m = qy^\ell$ for all $q \in Q$, and consequently, $\delta(q_0, z^m) \in F$ if and only if $\delta(q_0, y^\ell) \in F$ if and only if $\delta(q_0, y) \in F$. The equality $qy^\ell y^\ell = qy^\ell$ yields $qz = qy^\ell xy^\ell = qy^\ell y^\ell xy^\ell = qy^\ell z$, and therefore, $qz^m = qy^\ell z^m$. Since $Qy^\ell = Q'$, we have $qy^\ell z^m = qy^\ell$ and $qz^m = qy^\ell$ as required. Therefore, the message $(y^\ell xy^\ell)^m$ over C has two factorizations with distinct sets of code words contradicting the fact that C is SD. ■

The following is the SD version of a result of Lempel [1]. Here a code C is called a *prefix* (respectively, *suffix*) code if none of its words begins (respectively, ends) with a shorter word of C.

COROLLARY *1. No SD code contains* a *full prefix code or* a *full suffix code as* a *proper subcode.*

PROOF. The result follows from Theorem 2 and the fact that all prefix codes and all suffix codes are UD since there is no path from **open** to **close** for such codes. ■

References:

1. Lempel, On multiset decipherable codes, *IEEE Transactions on Information Theory* 32, 714-716 (1986).
2. F. Guzman, Decipherability of codes, *Journal of Pure and Applied Algebra* **141,** 13-35 (1999).
3. F. Guzman, A complete list of small proper MSD and SD codes (submitted).
4. F. Blanchet-Sadri, On unique, multiset, and set decipherability of three-word codes, *IEEE Transactions on Information Theory* (to appear).
5. F. Blanchet-Sadri, Combinatorics on three-word codes, *Discrete Applied Mathematics* (submitted).
6. F. Blanchet-Sadri and T. Howell, A note on decipherability of three-word codes, *RAIRO Informatique Theorique et Applications* (submitted).
7. R.G. Gallagher, *Information Theory and Reliable Communications,* Wiley, New York, (1968).

8. Restivo, A note on multiset decipherable codes, *IEEE Transactions on Information Theory* **35,** 662-663 (1989).
9. Apostolico and R. Giancarlo, Pattern matching machine implementation of a fast test for unique decipherability, *Information Processing Letters* **18,** 155-158 (1984).
10. T. Head and A. Weber, Deciding multiset decipherability, *IEEE Transactions on Information Theory* **41,** 291-297 (1995).
11. Hoffmann, A test on unique decipherability, In *MFCS 84, Lecture Notes in Computer Science, Volume 176,* pp. 50-63, Springer-Verlag, Berlin, (1984).
12. M. Rodeh, A fast test for unique decipherability based on suffix trees, *IEEE Transactions on Information Theory* **28,** 648-651 (1982).
13. J. Berstel and D. Perrin, *Theory of Codes,* Academic Press, Orlando, FL, (1985).