

# Correlations of Partial Words<sup>\*</sup>

## (Extended Abstract)

F. Blanchet-Sadri<sup>1</sup>, Joshua D. Gafni<sup>2</sup>, and Kevin H. Wilson<sup>3</sup>

<sup>1</sup> Department of Computer Science, University of North Carolina,  
P.O. Box 26170, Greensboro, NC 27402–6170, USA, [blanchet@uncg.edu](mailto:blanchet@uncg.edu)

<sup>2</sup> Department of Mathematics, University of Pennsylvania,  
Philadelphia, PA 19104–6395, USA, [jgafni@sas.upenn.edu](mailto:jgafni@sas.upenn.edu)

<sup>3</sup> Department of Mathematics, University of Michigan,  
Ann Arbor, MI 48109–1043, USA, [khwilson@umich.edu](mailto:khwilson@umich.edu)

**Abstract.** *Partial words* are strings over a finite alphabet that may contain a number of “do not know” symbols. In this paper, we introduce the notions of binary and ternary correlations, which are binary and ternary vectors indicating the periods and weak periods of partial words. Extending a result of Guibas and Odlyzko, we characterize precisely which of these vectors represent the (weak) period sets of partial words and prove that all valid correlations may be taken over the binary alphabet. We show that the sets of all such vectors of a given length form distributive lattices under inclusion. We also show that there is a well defined minimal set of generators for any binary correlation of length  $n$  and demonstrate that these generating sets are the primitive subsets of  $\{1, 2, \dots, n - 1\}$ . Finally, we investigate the number of correlations of length  $n$ .

## 1 Introduction

*Words*, sequences or strings of symbols from a finite alphabet, arise naturally in several areas of mathematical sciences. Notions and techniques related to periodic structures in words find applications in virtually every area of theoretical and applied computer science, notably in text processing, data compression, coding, computational biology, string searching and pattern matching algorithms. Repeated patterns and related phenomena in words have played over the years a central role in the development of combinatorics on words, and have been highly valuable tools for the design and analysis of algorithms.

The first significant results on periodicity are the theorem of Fine and Wilf [9] and the critical factorization theorem [7]. These two fundamental results refer to two kinds of phenomena concerning periodicity: The theorem of Fine and Wilf

---

<sup>\*</sup> This material is based upon work supported by the National Science Foundation under Grant No. DMS-0452020. A World Wide Web server interface has been established at [www.uncg.edu/mat/research/correlations](http://www.uncg.edu/mat/research/correlations) for automated use of the program. We thank the referees of a preliminary version of this paper for their very valuable comments and suggestions.

considers the simultaneous occurrence of different periods in one string, whereas the critical factorization theorem relates local and global periodicity of strings. Starting from these basic classical results, the study of periodicity has grown along both directions. Reference [15] contains a systematic and self-contained exposition of this theory, including more recent significant results such as an unexpected theorem of Guibas and Odlyzko which gives the structure of the set of periods of a string [10].

In many practical applications, such as DNA sequence analysis, repetitions admit a certain variation between copies of the repeated pattern because of errors due to mutation, experiments, etc. Approximate repeated patterns, or repetitions where errors are allowed, are playing a central role in different variants of string searching and pattern matching problems. *Partial words*, or strings that may have a number of “do not know” symbols (also called “holes”), have acquired great importance in this context [11,12,13,14,17]. Another application area of current interest for the study of partial words is data communication where some information may be missing, lost, or unknown. In their seminal and fundamental work [1], Berstel and Boasson introduced this notion of partial word and proved a theorem analogous to the periodicity theorem of Fine and Wilf for the one-hole case. After them, Blanchet-Sadri and Hegstrom extended this result to partial words with two and three holes [5], and finally Blanchet-Sadri extended it to arbitrary partial words [2]. Blanchet-Sadri and co-authors have developed this line of research of periodicity on partial words and obtained the first algorithms in the context of partial words. In particular, they extended the critical factorization theorem to partial words with an arbitrary number of holes [4,6] and Guibas and Odlyzko’s theorem to partial words with one hole [3].

In [10], Guibas and Odlyzko consider the period sets of strings of length  $n$  over a finite alphabet, and specific representations of them, *(auto)correlations*, which are binary vectors of length  $n$  indicating the periods. Among the possible  $2^n$  bit vectors, only a small subset are valid correlations. There, they provide characterizations of correlations, asymptotic bounds on their number, and a recurrence for the *population size* of a correlation, that is, the number of strings sharing a given correlation. In [16], Rivals and Rahmann show that there is redundancy in period sets and introduce the notion of an *irreducible* period set. They prove that  $\Gamma_n$ , the set of all correlations of length  $n$ , is a lattice under set inclusion and does not satisfy the Jordan-Dedekind condition. They propose the first efficient enumeration algorithm for  $\Gamma_n$  and improve upon the previously known asymptotic lower bounds on the cardinality of  $\Gamma_n$ . Finally, they provide a new recurrence to compute the number of strings sharing a given period set, and exhibit an algorithm to sample uniformly period sets through irreducible period sets.

In the case of partial words, there are two notions of periodicity: one is that of *period*, the other is that of *weak period*. In this paper, we study the combinatorics of possible sets of periods and weak periods of partial words in a similar way as it was done for the structure of all global periods of words. In Section 3, we introduce the notions of binary and ternary correlations, which are binary

and ternary vectors indicating the periods and weak periods of partial words. Extending the result of Guibas and Odlyzko, we characterize precisely which of these vectors represent the (weak) period sets of partial words and prove that all valid correlations may be taken over the binary alphabet. In Section 4, we show that the sets of all such vectors of a given length form distributive lattices under inclusion extending results of Rivals and Rahmann. We also show that there is a well defined minimal set of generators for any binary correlation of length  $n$  and demonstrate in Section 5 that these generating sets are the primitive subsets of  $\{1, 2, \dots, n-1\}$ . Finally, we investigate the number of correlations of length  $n$ .

## 2 Definitions, notations, and preliminary results

Traditionally, a (*full*) word  $u$  is defined as a function  $u : \{0, 1, \dots, n-1\} \rightarrow A$  for some  $n \geq 0$  and some nonempty, finite set  $A$ , called the *alphabet*. The *length*  $n$  is denoted  $|u|$  and sometimes the word is written explicitly as  $u = u(0)u(1) \cdots u(n-1)$ . When  $n = 0$  we say the word is *empty* and denote it by  $\varepsilon$ . We denote the set of all words of length  $n$  over the alphabet  $A$  by  $A^n$  and the set of all words over  $A$  by  $A^*$ .

A *partial word* is defined similarly except  $u$  is a *partial* function. We define  $D(u)$  to be the *domain of  $u$* , i.e., the set of  $i \in \{0, 1, \dots, n-1\}$  such that  $u(i)$  is defined. Moreover, we define the *companion of  $u$*  to be the full word  $u_\diamond : \{0, 1, \dots, n-1\} \rightarrow A \cup \{\diamond\}$  defined by  $u_\diamond(i) = u(i)$  if  $i \in D(u)$  and  $u_\diamond(i) = \diamond$  otherwise. Finally we define  $H(u) = \{0, 1, \dots, n-1\} \setminus D(u)$  to be the set of *holes of  $u$* . Throughout this paper  $u$  and  $u_\diamond$  will be used interchangeably. For example,  $u = abb\diamond bb\diamond a$  is a partial word where  $D(u) = \{0, 1, 2, 4, 5, 7\}$  and  $H(u) = \{3, 6\}$ . We say that  $A_\diamond^n$  is the set of partial words of length  $n$  over the alphabet  $A$  and that  $A_\diamond^*$  is the set of *all* partial words (including  $\varepsilon$ ) over the alphabet  $A$ .

Partial words allow for two weakenings of equality which we call containment and compatibility. We say that the partial word  $u$  is *contained in* the partial word  $v$  (denoted  $u \subset v$ ) provided that  $|u| = |v|$ ,  $D(u) \subseteq D(v)$  and for all  $i \in D(u)$  we have that  $u(i) = v(i)$ . As a weaker notion, we say that the partial words  $u$  and  $v$  are *compatible* (denoted  $u \uparrow v$ ) provided that there exists another partial word  $w$  such that  $u \subset w$  and  $v \subset w$ . An equivalent formulation of compatibility is that  $|u| = |v|$  and for all  $i \in D(u) \cap D(v)$  we have that  $u(i) = v(i)$ . As an example,  $u = abb\diamond bb\diamond a$  and  $v = \diamond bbab\diamond ba$  are compatible with  $w = abbabbba$ . For a partial word  $u$ , we denote by  $C(u)$  the set of all partial words compatible with  $u$ . More specifically,  $C(u) = \{v \mid u \uparrow v\}$ .

We say that a partial word  $u$  is (*strongly*)  $p$ -*periodic* provided that  $u(i) = u(j)$  for all  $i, j \in D(u)$  with  $i \equiv j \pmod{p}$ . Moreover, we say that all partial words have period 0. We denote the set of all periods of  $u$  which are less than  $|u|$  by  $\mathcal{P}(u)$ . Similarly we say that a partial word  $u$  is *weakly  $p$ -periodic* provided that whenever  $0 \leq i < n-p$  and  $i, i+p \in D(u)$  we have  $u(i) = u(i+p)$ . We call the set of weak periods of  $u$  which are less than  $|u|$  by  $\mathcal{P}'(u)$ . It is obvious that  $\mathcal{P}(u) \subseteq \mathcal{P}'(u)$  and in the case of full words,  $\mathcal{P}(u) = \mathcal{P}'(u)$  since  $D(u) = \{0, 1, \dots, |u|-1\}$ . In general this equality does not hold. As an example, consider the partial word  $ab\diamond bbb\diamond bbb$ ,

which is weakly 2-periodic but *not* 2-periodic. When  $q \in \mathcal{P}'(u) \setminus \mathcal{P}(u)$  we say that  $u$  has a *strictly* weak period of  $q$ . Note that if for some  $n$  we have that  $u, v \in A_\diamond^n$  and  $u \subset v$ , then  $\mathcal{P}(v) \subseteq \mathcal{P}(u)$  and  $\mathcal{P}'(v) \subseteq \mathcal{P}'(u)$ .

We will say that the *greatest lower bound* of a pair of partial words  $u$  and  $v$  of length  $n$  is the partial word  $u \wedge v$  with  $D(u \wedge v) = \{0 \leq i < n \mid i \in D(u) \cap D(v) \text{ and } u(i) = v(i)\}$  and  $(u \wedge v)(i) = u(i) = v(i)$  for all  $i \in D(u \wedge v)$ . Consider for example the partial words  $u = \text{abbbb} \diamond a$  and  $v = \diamond \text{bbab} \diamond ba$  where  $u \wedge v = \diamond \text{bb} \diamond b \diamond \diamond a$ . Note that  $u \wedge v$  is constructed so that  $(u \wedge v) \subset u$  and  $(u \wedge v) \subset v$ . Moreover, it is easily seen that  $u \wedge v$  is *maximal* in the sense that for all partial words  $w$  which satisfy  $w \subset u$  and  $w \subset v$  we have that  $w \subset (u \wedge v)$ . One property we notice immediately about the greatest lower bound is that if  $u, v \in A_\diamond^n$ , then  $\mathcal{P}(u) \cup \mathcal{P}(v) \subseteq \mathcal{P}(u \wedge v)$  and  $\mathcal{P}'(u) \cup \mathcal{P}'(v) \subseteq \mathcal{P}'(u \wedge v)$ .

For any  $0 \leq p < |u|$  and  $0 \leq i < p$ , define  $u_{i,p} = u(i)u(i+p)u(i+2p)\cdots$ , the  $i$ th  $p$ -word of  $u$ . Clearly,  $p \in \mathcal{P}(u)$  if and only if  $u_{i,p}$  is 1-periodic for all  $0 \leq i < p$ . Similarly,  $p \in \mathcal{P}'(u)$  if and only if  $u_{i,p}$  is weakly 1-periodic for all  $0 \leq i < p$ .

### 3 Characterizations of correlations

The major result of [10] was a complete characterization of the possible sets of periods for full words of arbitrary length. Guibas and Odlyzko stated their results not in terms of sets of periods but in terms of bit vectors which they called correlations. For a (full) word  $u$ , let  $v$  be the bit vector of length  $|u|$  for which  $v_i = 1$  whenever  $i \in \mathcal{P}(u)$  and  $v_i = 0$  otherwise. We call  $v$  the *correlation* of  $u$ . For example, the word  $\text{abbababbab}$  has periods 5 and 8 and thus has correlation  $1000010010$ . This representation gave them a useful method of representing sets of periods in concise ways and allowed them to prove the main result of their paper. We now recall their theorem, for which we will need a definition.

**Definition 1.** A bit vector  $v$  of length  $n$  is said to satisfy the

- Forward propagation rule provided that for all  $0 \leq p < q < n$  such that  $v_p = v_q = 1$  we have that  $v_{p+i(q-p)} = 1$  for all integers  $i$  satisfying  $2 \leq i < (n-p)/(q-p)$ ,
- Backward propagation rule provided that for any nonnegative integers  $p$  and  $q$  less than  $n$  such that  $0 \leq p < q < 2p$  and  $v_p = v_q = 1$  and  $v_{2p-q} = 0$  we have that  $v_{p-i(q-p)} = 0$  for all  $i = 2, \dots, \min\{p/(q-p), (n-p)/(q-p)\}$ .

**Theorem 1. Guibas and Odlyzko** [10] For correlation  $v$  of length  $n$  the following are equivalent:

1. There exists a word over the binary alphabet with correlation  $v$ .
2. There exists a word over some alphabet with correlation  $v$ .
3. The correlation  $v$  satisfies the forward and backward propagation rules.

**Corollary 1.** For any alphabet  $A$  and any word  $u \in A^*$ , there exists a word  $v \in \{a, b\}^*$  such that  $\mathcal{P}(v) = \mathcal{P}(u)$ .

In this section, we follow the example of Guibas and Odlyzko and completely characterize the possible sets of periods and weak periods of partial words. To do so we first extend their definition of a “correlation” to incorporate the difference between strictly weak periods and strong periods, a difference which does not occur in the case of full words.

**Definition 2.** *Let  $P$  and  $Q$  be sets. We say that the pair  $(P, Q)$  is a ternary correlation of length  $n$  provided that there exists a partial word  $u \in A_\diamond^n$  such that  $P = \mathcal{P}(u)$  and  $Q = \mathcal{P}'(u) \setminus \mathcal{P}(u)$ . Such a pair we will denote by  $P/Q$ . For a given ternary correlation  $P/Q$  of length  $n$ , we define its correlation vector  $v$  to be the ternary vector for which  $v_i = 1$  whenever  $i \in P$ ,  $v_i = 2$  whenever  $i \in Q$ , and  $v_i = 0$  otherwise. We will say that  $\mathcal{P}(v) = \{0 \leq i < n \mid v_i = 1\}$  and  $\mathcal{P}'(v) = \{0 \leq i < n \mid v_i > 0\}$ . When  $Q = \emptyset$ , we will call the correlation  $P/Q$  a binary correlation.*

We begin the process of characterizing the correlations of partial words by recording two facts. (1) The first formalizes a relatively obvious property of the periods of full words: For all integers  $p \geq 0$  define  $\langle p \rangle_n$  to be the set of nonnegative integers less than  $n$  which are multiples of  $p$ . Then for all  $u \in A^n$ ,

$$\mathcal{P}(u) = \bigcup_{p \in P} \langle p \rangle_n$$

for some  $P \subseteq \{0, 1, 2, \dots, n-1\}$ . (2) The second characterizes the relationship between partial words and the words which are compatible with them: If  $u$  is a partial word over an alphabet  $A$ , then

$$\mathcal{P}(u) = \bigcup_{w \in C(u) \cap A^*} \mathcal{P}(w)$$

For example, consider the partial word  $u = abca \diamond cabca$  over the alphabet  $A = \{a, b, c\}$ . Then  $\mathcal{P}(u) = \{3, 6, 9, 10\} = \mathcal{P}(w_1) \cup \mathcal{P}(w_2) \cup \mathcal{P}(w_3)$  where  $w_1 = abcaacabca$ ,  $w_2 = abcababca$  and  $w_3 = abcaccabca$  are the words  $w$  satisfying  $w \in C(u) \cap A^*$ .

We are now ready to state the first part of our characterization theorem. This part of the theorem completely characterizes the set of binary correlations of all partial words. In the sequel, we only use the subscript  $n$  in  $\langle p \rangle_n$  when its value is not clear from context.

**Theorem 2.** *For any finite collection  $u_1, u_2, \dots, u_k$  of full words of length  $n$  over an alphabet  $A$ , there exists a partial word  $w$  of length  $n$  over the alphabet  $\{a, b\}$  with  $\mathcal{P}(w) = \mathcal{P}'(w) = \mathcal{P}(u_1) \cup \mathcal{P}(u_2) \cup \dots \cup \mathcal{P}(u_k)$ .*

*Proof. (Sketch)* The case  $k = 1$  follows from Corollary 1. Moreover, since  $\varepsilon$  is the only word of length 0, the case  $n = 0$  forces  $k = 1$  and so we assume that  $k > 1$  and  $n > 0$ . Then from (1) we have that

$$\bigcup_{j=1}^k \mathcal{P}(u_j) = \bigcup_{p \in P} \langle p \rangle$$

for some  $P \subseteq \{0, 1, \dots, n-1\}$ . Thus for all  $1 \leq j \leq k$ , we assume that  $\mathcal{P}(u_j) = \langle p_j \rangle$  for some  $0 \leq p_j < n$ .

With these assumptions, we move on to the case when  $k = 2$ . For notational clarity we set  $u = u_1$ ,  $v = u_2$ ,  $\mathcal{P}(u) = \langle p \rangle$ , and  $\mathcal{P}(v) = \langle q \rangle$  for some  $0 \leq p < q < n$ . Define

$$\omega_p = \begin{cases} ab^{n-1} & \text{if } p = 0 \\ (ab^{p-1})^k ab^{r-1} & \text{if } p > 0 \text{ and } r > 0 \\ (ab^{p-1})^k & \text{otherwise} \end{cases}$$

where  $n = kp + r$  with  $0 \leq r < p$ . Similarly define  $\omega_q$ . Obviously  $\mathcal{P}(\omega_p) = \langle p \rangle$  and  $\mathcal{P}(\omega_q) = \langle q \rangle$ . We claim that  $\mathcal{P}(\omega_p \wedge \omega_q) = \mathcal{P}(\omega_p) \cup \mathcal{P}(\omega_q)$ .

Moreover, we see that  $\omega_p \wedge \omega_q$  has no strictly weak periods. Assume the contrary and let  $\xi \in \mathcal{P}'(\omega_p \wedge \omega_q) \setminus \mathcal{P}(\omega_p \wedge \omega_q)$ . Then there exist  $i, j \in D(\omega_p \wedge \omega_q)$  such that  $i \equiv j \pmod{\xi}$  and  $(\omega_p \wedge \omega_q)(i) = a$  and  $(\omega_p \wedge \omega_q)(j) = b$ , and for all  $0 \leq k < n$  such that  $k \equiv i \pmod{\xi}$  and  $k$  is strictly between  $i$  and  $j$  we have  $k \in H(\omega_p \wedge \omega_q)$ . Let  $k$  be such that  $|i - k|$  is minimized (that is, if  $i < j$  then  $k$  is minimal and if  $i > j$  then  $k$  is maximal). This minimal distance is obviously  $\xi$ . Then  $p$  and  $q$  divide  $i$  and at least one of them divides  $k$ . But we see that only one of  $p$  and  $q$  divides  $k$ , for if both did then  $(\omega_p \wedge \omega_q)(k) = a \neq \diamond$ . Without loss of generality let  $p|k$ . But as  $p|i$  and  $p|k$ , we have  $p||i - k| = \xi$ . Then since  $\omega_p$  is  $p$ -periodic, we have that  $\omega_p(l) = \omega_p(i) = a$  for all  $l \equiv i \pmod{p}$ . But  $j \equiv i \pmod{\xi}$  and  $p|\xi$ , so  $j \equiv i \pmod{p}$ . Therefore,  $\omega_p(j) = a$  and thus  $(\omega_p \wedge \omega_q)(j) \neq b$ , a contradiction. Now let  $k > 2$  and let  $\{p_1, \dots, p_k\} \subseteq \{0, 1, \dots, n-1\}$  be the periods such that  $\mathcal{P}(u_j) = \langle p_j \rangle$ . We claim that  $\mathcal{P}(\omega_{p_1} \wedge \dots \wedge \omega_{p_k}) = \mathcal{P}(\omega_{p_1}) \cup \dots \cup \mathcal{P}(\omega_{p_k})$  and that  $\omega_{p_1} \wedge \dots \wedge \omega_{p_k}$  has no strictly weak periods.  $\square$

Theorem 2 tells us that every union of possible correlations of full words over any alphabet is the correlation of a binary partial word. But (2) tells us that the period set of every partial word over any alphabet (including the binary alphabet) is the union of the period sets of all full words compatible with it. Thus, we have a bijection between these sets which we record as the following corollary.

**Corollary 2.** *The set of valid binary correlations  $P/\emptyset$  of length  $n$  over the binary alphabet is precisely the set of unions of valid correlations of full words of length  $n$  over all nonempty alphabets.*

In light of (2), the following corollary is essentially a rephrasing of the previous corollary. But as a concept, this corollary is important enough to deserve special attention.

**Corollary 3.** *The set of valid binary correlations  $P/\emptyset$  over an alphabet  $A$  with  $|A| \geq 2$  is the set of valid binary correlations over the binary alphabet. Phrased differently, if  $u$  is a partial word over an alphabet  $A$ , then there exists a binary partial word  $v$  such that  $\mathcal{P}(v) = \mathcal{P}(u)$ .*

Theorem 2 and Corollaries 2 and 3 give us three equivalent characterizations of valid binary correlations of partial words over an arbitrary alphabet. They

do not mention at all, though, the effect of strictly weak periods. The following theorem shows that the characterization is actually rather elegant.

**Theorem 3.** *A ternary correlation  $P/Q$  of length  $n$  is valid if and only if*

1.  $P$  is the nonempty union of sets of the form  $\langle p \rangle_n$ ,
2. For each  $q \in Q$ , there exists an integer  $2 \leq m < \frac{n}{q}$  such that  $mq \notin P \cup Q$ .

*Proof. (Sketch)* First, if  $Q = \emptyset$  then we are in the case of Corollaries 2 and 3. Thus we consider only the case when  $Q \neq \emptyset$ . We begin by taking a triple  $(P, Q, n)$  satisfying the above conditions along with the assumption that  $n$  is at least 3 since the cases of zero-letter, one-letter, and two-letter partial words are trivial by simple enumeration considering all possible renamings of letters. So we may now define

$$\psi_Q = \bigwedge_{q \in Q} \psi_q \quad \omega_P = \bigwedge_{p \in P} \omega_p$$

where  $\psi_q = ab^{q-1} \diamond b^{n-q-1}$  with  $1 \leq q < n$ ,  $a, b \in A$  are distinct letters, and  $\omega_p$  is as in the proof of Theorem 2. Notice that  $0 \notin Q$  since  $0 \in P$  and then  $0m \in P$  for all integers  $m$ . Thus,  $\psi_Q$  is well-defined. Then we claim that  $u = \omega_P \wedge \psi_Q$  is a partial word with correlation  $P/Q$ .

We claim that  $P = \mathcal{P}(u)$  and since  $P \cup Q \subseteq \mathcal{P}'(u)$  it suffices to show that if  $q \in \mathcal{P}'(u) \setminus \mathcal{P}(u)$  then  $q \in Q$ . Since  $q \in \mathcal{P}'(u) \setminus \mathcal{P}(u)$  we have that some  $u_{i,q}$  contains both  $a$  and  $b$ . But the only possible location of  $a$  is 0, so we may write this as  $u(0) = a$ ,  $u(qj) = \diamond$ , and  $u(qk) = b$  for some  $k \geq 2$  and  $0 < j < k$ . But notice then that  $u$  does not have period  $q$  so  $q \notin P$ . Thus, since  $u(q) = \diamond$ , we have that  $q \in Q$  and have thus completed this direction of the theorem. Now consider the other direction, i.e., if we are given a partial word  $u$  with correlation  $P/Q$ , then  $P/Q$  satisfies our conditions. By Theorem 2 we have that the first condition must be met and we claim that the second condition must be met as well.  $\square$

In analogy to Corollary 3, we record the following fact.

**Corollary 4.** *The set of valid ternary correlations  $P/Q$  over an alphabet  $A$  with  $|A| \geq 2$  is the same as the set of valid ternary correlations over the binary alphabet. Phrased differently, if  $u$  is a partial word over an alphabet  $A$ , then there exists a binary partial word  $v$  with  $\mathcal{P}(v) = \mathcal{P}(u)$  and  $\mathcal{P}'(v) = \mathcal{P}'(u)$ .*

We end this section with some consequences of Theorem 3. Having completely characterized the sets of binary and ternary correlations of partial words and having shown that all valid binary and ternary correlations may be taken as over the binary alphabet, we give these sets names. In the sequel we shall let  $\Delta_n$  be the set of all valid binary correlations of partial words of length  $n$  and  $\Delta'_n$  the set of valid ternary correlations of length  $n$ .

As a first consequence of Theorem 3 we notice that for a given ternary correlation  $v \in \Delta'_n$ , we have that  $v_i \neq 2$  for all  $i > \lfloor \frac{n-1}{2} \rfloor$ . Another consequence of Theorem 3 is that for all  $v \in \Delta_n$  we have that

$$\mathcal{P}(v) = \bigcup_{p \in P} \langle p \rangle_n$$

for some  $P \subseteq \{0, 1, \dots, n-1\}$ , and we say that  $P$  *generates* the correlation  $v$ . One such  $P$  is  $\mathcal{P}(v)$ . But in general there are strictly smaller  $P$  which have this property. For example, if  $v = 1001001101$  then  $\mathcal{P}(v) = \{0, 3, 6, 7, 9\}$ . While  $P = \mathcal{P}(v)$  will generate this set, we see that  $P = \{0, 3, 6, 7\}$ ,  $\{3, 6, 7\}$ ,  $\{0, 3, 7\}$ , or  $\{3, 7\}$  (among others) will as well. On the other hand, we see that there is a well defined *minimal* set of generators. That is, for every  $v \in \Delta_n$  there is a set  $R(v)$  such that for any set  $P$  which generates  $v$  we have that  $R(v) \subseteq P$ . Namely, this is the set of nonzero  $p \in \mathcal{P}(v)$  such that for all  $q \in \mathcal{P}(v)$  with  $p \neq q$  we have that  $q \not\mid p$ . For if there is  $q$  distinct from  $p$  such that  $q \mid p$  then we have that all multiples of  $p$  are also multiples of  $q$ , i.e.,  $\langle p \rangle \subseteq \langle q \rangle$ . Moreover, we see since there are no divisors of the elements of  $R(v)$  in  $\mathcal{P}(v)$  that the only  $p \in \mathcal{P}(v)$  which can generate  $r \in R(v)$  is  $r$  itself. Thus we have achieved minimality.

We will call  $R(v)$  the *irreducible period set of  $v$* . For partial words of length  $n$ , we define  $\Phi_n$  to be the set of all irreducible period sets. Moreover, we see that there is an obvious bijective correspondence between  $\Phi_n$  and  $\Delta_n$  given by the function  $R : \Delta_n \rightarrow \Phi_n$  in one direction and its inverse  $E : \Phi_n \rightarrow \Delta_n$  defined as

$$E(P) = \bigcup_{p \in P} \langle p \rangle_n$$

#### 4 Structural properties of $\Delta_n$ , $\Delta'_n$ and $\Phi_n$

In [16] Rivals and Rahmann defined the set of all valid correlations of full words of length  $n$  as  $\Gamma_n$ . They then defined a notion of irreducible period set based on forward propagation. Specifically, they noticed that (like partial words), some periods are implied by other periods because of the forward propagation rule. An example is that if a twelve-letter word has periods 7 and 9 then it must also have period 11 since  $11 = 7 + 2(9 - 7)$ . They then gave for any  $v \in \Gamma_n$ , conditions for a period set to be an irreducible period set associated with  $v$  and showed that this minimal set of periods exists and is unique. In the above example,  $\{0, 7, 9, 11\}$  would correspond to  $\{0, 7, 9\}$ . The set of these irreducible period sets they called  $A_n$ .

Our notion of irreducible periods and Rivals and Rahmann's differ in a fundamental way. Specifically, their definition relied on forward propagation. This rule does not hold in the case of partial words. For example, the proof of Theorem 3 tells us that  $abbbbbbb \diamond b \diamond bb$  has periods 7 and 9 but does *not* have period 11. Thus,  $\{7, 9, 11\}$  is irreducible in the sense of partial words, but not in the sense of full words.

The idea of reduction is still present though. And in [16] Rivals and Rahmann went on to show several structural properties of  $\Gamma_n$  and  $A_n$ . Specifically, they showed that  $\Gamma_n$  is a lattice under inclusion which does not satisfy the Jordan-Dedekind condition, a criterion which stipulates that all maximal chains between two elements of a poset are of equal length. Violating this condition implies that  $\Gamma_n$  is neither distributive, modular, nor a matroid. They also showed that while  $\Gamma_n$  is not a lattice that it *does* satisfy the Jordan-Dedekind condition as a poset.

Because of the analogies between  $\Gamma_n$  and  $\Delta_n$  and  $\Delta'_n$  as well as the analogies between  $\Lambda_n$  and  $\Phi_n$ , we now investigate the structural properties of  $\Delta_n$ ,  $\Delta'_n$  and  $\Phi_n$ . In order to highlight the differences between the cases of full words and partial words, the structure of this section closely follows the structure of the analogous section of [16]. In particular we show that both  $\Delta_n$  and  $\Delta'_n$  are distributive lattices under inclusion (suitably defined in the case of  $\Delta'_n$ ). On the other hand, we show that  $\Phi_n$  is not a lattice but does satisfy the Jordan-Dedekind condition.

First, we blur the lines between the correlation *vector*  $v \in \Delta_n$  and the associated set of periods  $\mathcal{P}(v)$ . Specifically, we say that for any  $u, v \in \Delta_n$  we have that  $u \subseteq v$  if and only if  $\mathcal{P}(u) \subseteq \mathcal{P}(v)$  and  $p \in u$  if and only if  $p \in \mathcal{P}(u)$ . Moreover, we define  $u \cap v$  and  $u \cup v$  to be the unique vectors with  $\mathcal{P}(u \cap v) = \mathcal{P}(u) \cap \mathcal{P}(v)$  and  $\mathcal{P}(u \cup v) = \mathcal{P}(u) \cup \mathcal{P}(v)$ . It is easy to see that if  $u, v \in \Delta_n$  then  $u \cap v \in \Delta_n$  and  $u \cup v \in \Delta_n$ . Moreover, the pair  $(\Delta_n, \subseteq)$  is a poset with a null element and a universal element. Namely the null element is  $10^{n-1}$  and the universal element is  $1^n$ . One of the theorems of [16] is that the set of correlations of full words form a lattice that does not satisfy the Jordan-Dedekind condition. Thus it is neither distributive nor modular. But since the meet and the join of binary correlations are the *set intersection* and *set union* of the correlations, we have the following theorem.

**Theorem 4.** *The poset  $(\Delta_n, \subseteq)$  is a distributive lattice and thus satisfies the Jordan-Dedekind condition.*

Second, we expand our considerations to  $\Delta'_n$ , the set of ternary correlations of partial words of length  $n$ , and show that  $\Delta'_n$  is a lattice again with respect to inclusion, which we define suitably. Consider ternary correlations  $u, v \in \Delta'_n$ . We define the intersection of  $u$  and  $v$  as the ternary vector  $u \cap v$  such that  $\mathcal{P}(u \cap v) = \mathcal{P}(u) \cap \mathcal{P}(v)$  and  $\mathcal{P}'(u \cap v) = \mathcal{P}'(u) \cap \mathcal{P}'(v)$ . Equivalently we might say that  $(u \cap v)_i = 0$  if either  $u_i = 0$  or  $v_i = 0$ , 1 if  $u_i = v_i = 1$ , and 2 otherwise. Note that  $\Delta'_n$  is closed under intersection.

We may define the union in the analogous way, specifically, for  $u, v \in \Delta'_n$  we say that  $\mathcal{P}(u \cup v) = \mathcal{P}(u) \cup \mathcal{P}(v)$  and that  $\mathcal{P}'(u \cup v) = \mathcal{P}'(u) \cup \mathcal{P}'(v)$ . Equivalently,  $u \cup v$  is the ternary vector satisfying  $(u \cup v)_i = 0$  if  $u_i = v_i = 0$ , 1 if either  $u_i = 1$  or  $v_i = 1$ , and 2 otherwise. Unlike unions of binary correlations, the union of two ternary correlations is not necessarily again a ternary correlation. For example, consider the correlations  $u = 102000101$  and  $v = 100010001$ . The union of these two correlations is  $u \cup v = 102010101$ , which violates the second condition of Theorem 3. Specifically, there is no  $q \geq 2$  such that  $(u \cup v)_{2q} = 0$ . Finally, for  $u, v \in \Delta'_n$  we say that  $u \subseteq v$  provided that  $\mathcal{P}(u) \subseteq \mathcal{P}(v)$  and  $\mathcal{P}'(u) \subseteq \mathcal{P}'(v)$ . Equivalently we may say that  $u \subseteq v$  provided that whenever  $u_i > 0$  we have that  $u_i \geq v_i > 0$ . Or more explicitly,  $u \subseteq v$  provided that whenever  $u_i = 1$  that  $v_i = 1$  and whenever  $u_i = 2$  that  $v_i = 1$  or  $v_i = 2$ . Under these definitions, the pair  $(\Delta'_n, \subseteq)$  is a poset with null element  $10^{n-1}$  and universal element  $1^n$ .

**Theorem 5.** *The poset  $(\Delta'_n, \subseteq)$  is a lattice.*

*Proof. (Sketch)* First,  $\Delta'_n$  is closed under intersection. Second, the pair  $(\Delta'_n, \subseteq)$  is a poset. Now, we do not have the union of the two correlations to explicitly define the join. One method of proving that the join exists is to notice that the join of  $u, v \in \Delta'_n$  is the intersection of all elements of  $\Delta'_n$  which contain  $u$  and  $v$ . This intersection is guaranteed to be nonempty since  $\Delta'_n$  contains a universal element. On the other hand, we can modify the union slightly such that we obtain the join constructively. Consider the example above in which  $u = 102000101$  and  $v = 100010001$  and  $u \cup v = 102010101$ . If we simply change  $(u \cup v)_2$  from 2 to 1, then we will have created a valid ternary correlation. Calling this vector  $u \vee v$  we see that  $u \subseteq u \vee v$  and that  $v \subseteq u \vee v$ . Thus, we generalize this operator by defining  $u \vee v$  to be the unique correlation satisfying  $\mathcal{P}'(u \vee v) = \mathcal{P}'(u) \cup \mathcal{P}'(v)$  and  $\mathcal{P}(u \vee v) = \mathcal{P}(u) \cup \mathcal{P}(v) \cup B(u \cup v)$  where  $B(u \cup v)$  is the set of all  $0 \leq q < n$  such that  $(u \cup v)_q = 2$  and there exists no  $k \geq 2$  such that  $(u \cup v)_{kq} = 0$ . That is,  $B(u \cup v)$  is the set of positions in  $u \cup v$  which do not satisfy the second condition of Theorem 3.

We claim that  $u \vee v$  is the unique join of  $u$  and  $v$  (and thus justify our use of the traditional notation  $\vee$  for our binary operation). Notice first that since  $\mathcal{P}(u \cup v) = \mathcal{P}(u) \cup \mathcal{P}(v)$  and  $\mathcal{P}'(u \cup v) = \mathcal{P}'(u) \cup \mathcal{P}'(v)$  that  $u \cup v \subseteq u \vee v$ . Thus we have that  $u \subseteq u \cup v \subseteq u \vee v$  and that  $v \subseteq u \cup v \subseteq u \vee v$ . We also see that  $u \vee v \in \Delta'_n$ . This follows from the fact that if  $p \in \mathcal{P}(u \vee v)$  then either  $p \in \mathcal{P}(u) \cup \mathcal{P}(v)$  or for all  $k \geq 1$  we have that  $kp \in \mathcal{P}'(u) \cup \mathcal{P}'(v)$ . In the first case, we then have that  $\langle p \rangle \subseteq \mathcal{P}(u) \cup \mathcal{P}(v) \subseteq \mathcal{P}(u \vee v)$ . In the second case, we see that all multiples of  $p$  are in  $\mathcal{P}'(u) \cup \mathcal{P}'(v)$ . Therefore, by the definition of  $u \vee v$  and the fact that the multiples of all multiples of  $p$  are again multiples of  $p$ , we must have that  $\langle p \rangle \subseteq \mathcal{P}(u \vee v)$ . Thus, using the  $\vee$  operator instead of the  $\cup$  operator resolves all conflicts with Theorem 3 and so  $u \vee v \in \Delta'_n$ . From here it suffices to show that it is minimal.  $\square$

Strangely, even though the join operation of  $\Delta'_n$  is more complicated than the join operation of  $\Delta_n$ , we still have that  $\Delta'_n$  is distributive and thus satisfies the Jordan-Dedekind condition. This is stated in the following theorem.

**Theorem 6.** *The lattice  $(\Delta'_n, \subseteq)$  is distributive and thus satisfies the Jordan-Dedekind condition.*

So unlike the lattice of correlations of full words which does not even satisfy the Jordan-Dedekind condition, the lattices of both binary and ternary correlations of partial words are distributive.

Finally we turn our attention to  $\Phi_n = R(\Delta_n)$ , the set of irreducible period sets of length  $n$ . For  $n \geq 3$ , we see immediately that the poset  $(\Phi_n, \subseteq)$  is *not* a join-semilattice since the sets  $\{1\}$  and  $\{2\}$  will never have a join since  $\{1\}$  is always maximal. On the other hand, we have that  $(\Phi_n, \subseteq)$  is a meet-semilattice as it contains a null element  $\emptyset$ . The meet of two elements of  $\Phi_n$  is simply their set theoretic intersection.

**Proposition 1.**  *$\Phi_n$  satisfies the Jordan-Dedekind condition.*

Notice that while there is a natural bijection between the lattice  $\Delta_n$  and the meet-semilattice  $\Phi_n$  given above by the maps  $R$  and  $E$ , we see immediately that these maps are not morphisms. For example, consider the period sets  $\{0, 1, 2, 3, 4\}$  and  $\{0, 2, 4\}$ . Then we see that  $\{0, 1, 2, 3, 4\} \cap \{0, 2, 4\} = \{0, 2, 4\}$  for which the corresponding irreducible period set is  $\{2\}$ . But  $R(\{0, 1, 2, 3, 4\}) = \{1\}$  and  $R(\{0, 2, 4\}) = \{2\}$ , a pair of irreducible period sets whose intersection is  $\emptyset \neq \{2\}$ .

## 5 Counting correlations

In this section we look at the number of correlations of partial words of a given length. In the case of binary correlations, we give bounds and link the problem to one in number theory, and in the case of ternary correlations we give an exact count.

A *primitive set* of integers is a subset  $S \subseteq \mathbb{N} = \{1, 2, \dots\}$  such that for any two distinct elements  $s, s' \in S$  we have that neither  $s$  divides  $s'$  nor  $s'$  divides  $s$ . We denote by  $\mathbb{P}_n$  the set of finite primitive sets of integers at most  $n$ . As  $\Phi_n$  and  $\mathbb{P}_{n-1}$  coincide, we have the relation  $\|\Delta_n\| = \|\Phi_n\| = \|\mathbb{P}_{n-1}\|$ . So if we can count the number of finite primitive sets of integers less than  $n$  then we can count the number of binary correlations of partial words of length  $n$ . We present some results on approximating this number.

**Theorem 7. Erdős** [8] *Let  $S$  be a finite primitive set of size  $k$  with elements less than  $n$ . Then  $k \leq \lfloor \frac{n}{2} \rfloor$ . Moreover, this bound is sharp.*

This bound tells us that the number of primitive sets of integers with elements less than  $n$  is *at most* the number of subsets of  $\{1, 2, \dots, n-1\}$  of size at most  $\lfloor \frac{n}{2} \rfloor$ . Moreover, the sharpness of the bound gives us that  $\|\Phi_n\| \geq 2^{\lfloor n/2 \rfloor}$ . Thus we have that

$$\frac{\ln 2}{2} \leq \frac{\ln \|\Phi_n\|}{n} \leq \ln 2$$

In [10], Guibas and Odlyzko showed that as  $n \rightarrow \infty$

$$\frac{1}{2 \ln 2} + o(1) \leq \frac{\ln \|\Gamma_n\|}{(\ln n)^2} \leq \frac{1}{2 \ln(3/2)} + o(1)$$

and in [16] Rivals and Rahmann improved the lower bound to

$$\frac{\ln \|\Gamma_n\|}{(\ln n)^2} \geq \frac{1}{2 \ln 2} \left(1 - \frac{\ln \ln n}{\ln n}\right)^2 + \frac{0.4139}{\ln n} - \frac{1.47123 \ln \ln n}{(\ln n)^2} + O\left(\frac{1}{(\ln n)^2}\right)$$

where  $\Gamma_n$  is the set of all valid correlations of full words. Thus the bounds we give, which show explicitly that  $\ln \|\Delta_n\| = \Theta(n)$ , demonstrate that the number of valid binary correlations of partial words is *much* greater than the number of valid correlations of full words.

We now show that the set of ternary correlations is actually much more tractable to count than the set of binary correlations. We first note two interesting

consequences of Theorem 3: (1) Let  $u$  be a partial word of length  $n$  and let  $p \in \mathcal{P}'(u)$ . Then  $p \in \mathcal{P}(u)$  if and only if  $kp \in \mathcal{P}'(u)$  for all  $0 \leq k < n/p$ . That is, a weak period is a strong period if and only if all of its multiples are also weak periods. (2) If  $S \subseteq \{1, 2, \dots, n-1\}$ , then there is a unique ternary correlation  $v \in \Delta'_n$  such that  $\mathcal{P}'(v) = S \cup \{0\}$ . We note that (2) agrees with the definition of the join forced upon us in Section 4. Considering all periods as *weak* periods and then determining which ones are actually strong periods is how we defined that operation. We note that (2) tells us as well that the cardinality of the set of ternary correlations is the same as the cardinality of the power set of  $\{1, 2, \dots, n-1\}$ . And thus the equality  $\|\Delta'_n\| = 2^{n-1}$  holds.

## References

1. Berstel, J., Boasson, L.: Partial Words and a Theorem of Fine and Wilf. *Theoret. Comput. Sci.* **218** (1999) 135–141
2. Blanchet-Sadri, F.: Periodicity on Partial Words. *Comput. Math. Appl.* **47** (2004) 71–82
3. Blanchet-Sadri, F., Chriscoe, Ajay: Local Periods and Binary Partial Words: An Algorithm. *Theoret. Comput. Sci.* **314** (2004) 189–216 [www.uncg.edu/mat/AlgBin](http://www.uncg.edu/mat/AlgBin)
4. Blanchet-Sadri, F., Duncan, S.: Partial Words and the Critical Factorization Theorem. *J. Combin. Theory Ser. A* **109** (2005) 221–245 [www.uncg.edu/mat/cft](http://www.uncg.edu/mat/cft)
5. Blanchet-Sadri, F., Hegstrom, Robert A.: Partial Words and a Theorem of Fine and Wilf Revisited. *Theoret. Comput. Sci.* **270** (2002) 401–419
6. Blanchet-Sadri, F., Wetzler, N.D.: Partial Words and the Critical Factorization Theorem Revisited. [www.uncg.edu/mat/research/cft2](http://www.uncg.edu/mat/research/cft2)
7. Césari, Y., Vincent, M.: Une Caractérisation des Mots Périodiques. *C.R. Acad. Sci. Paris* **268** (1978) 1175–1177
8. Erdős, P.: Note on Sequences of Integers No One of Which is Divisible by Another. *J. London Math. Soc.* **10** (1935) 126–128
9. Fine, N.J., Wilf, H.S.: Uniqueness Theorems for Periodic Functions. *Proc. Amer. Math. Soc.* **16** (1965) 109–114
10. Guibas, L.J., Odlyzko, A.M.: Periods in Strings. *J. Combin. Theory Ser. A* **30** (1981) 19–42
11. Kolpakov, R., Kucherov, G.: Finding Approximate Repetitions Under Hamming Distance. *Lecture Notes in Comput. Sci.* Vol. 2161. Springer-Verlag, Berlin (2001) 170–181
12. Kolpakov, R., Kucherov, G.: Finding Approximate Repetitions Under Hamming Distance. *Theoret. Comput. Sci.* **33** (2003) 135–156
13. Landau, G., Schmidt, J.: An Algorithm for Approximate Tandem Repeats. *Lecture Notes in Comput. Sci.* Vol. 684. Springer-Verlag, Berlin (1993) 120–133
14. Landau, G.M., Schmidt, J.P., Sokol, D.: An Algorithm for Approximate Tandem Repeats. *J. Comput. Biology* **8** (2001) 1–18
15. Lothaire, M.: *Algebraic Combinatorics on Words*. Cambridge University Press, Cambridge (2002)
16. Rivals, E., Rahmann, S.: Combinatorics of Periods in Strings. *J. Combin. Theory Ser. A* **104** (2003) 95–113
17. Schmidt, J.P.: All Highest Scoring Paths in Weighted Grid Graphs and Their Application to Finding All Approximate Repeats in Strings. *SIAM J. Comput.* **27** (1998) 972–992