ARTICLE IN PRESS

Discrete Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

On pseudorandomness of families of binary sequences

András Sárközy

Eötvös Loránd University, Department of Algebra and Number Theory, H-1117 Budapest, Pázmány Péter sétány 1/C, Hungary

ARTICLE INFO

Article history:
Received 18 February 2015
Received in revised form 28 July 2015
Accepted 31 July 2015
Available online xxxx

Dedicated to the memory of Levon H. Khachatrian

Keywords:
Pseudorandom
Family of binary sequences
Family complexity
Collision
Distance minimum
Avalanche effect
Cross-correlation measure

ABSTRACT

In cryptography one needs large families of binary sequences with strong pseudorandom properties. In the last decades many families of this type have been constructed. However, in many applications it is not enough if our family of "good" sequences is large, it is more important to know that it has a rich, complex structure, and the sequences in the family are "independent", and they are "far apart". Thus various measures have been introduced and applied for studying pseudorandomness of families of binary sequences: family complexity, collision, distance minimum, avalanche effect and cross-correlation measure. In this paper a survey of all these definitions and results will be presented.

1. The measures of pseudorandomness of binary sequences

Finite binary sequences with strong pseudorandom (briefly PR) properties play a crucial role in cryptography, in particular, they can be used as *key* sequence in the Vernam cipher. In order to decide whether a certain sequence can be used for such a purpose one needs quantitative measures for pseudorandomness of binary sequences. A classical measure of this type is the *linear complexity*. However, it measures only one rather special (although important) property, and if we want to be sure that the sequence can be used securely then we also have to check other properties. Thus in 1997 Mauduit and Sárközy [28] (and later others) introduced further *PR measures*. Here we will need only the two most important measures of this type:

Consider the binary sequence

$$E_N = (e_1, e_2, \dots, e_N) \in \{-1, +1\}^N.$$

Then the *well-distribution measure* of E_N is defined as

$$W(E_N) = \max_{a,b,t} \left| \sum_{j=0}^{t-1} e_{a+jb} \right|$$

where the maximum is taken over all a, b, t such that $a, b, t \in \mathbb{N}$ and $1 \leqslant a < a + (t-1)b \leqslant N$, while the *correlation measure of order* k of E_N is defined as

$$C_k(E_N) = \max_{M,D} \left| \sum_{n=1}^{M} e_{n+d_1} e_{n+d_2} \dots e_{n+d_k} \right|$$

E-mail address: sarkozy@cs.elte.hu.

http://dx.doi.org/10.1016/j.dam.2015.07.031
0166-218X/© 2015 Elsevier B.V. All rights reserved.

where the maximum is taken over all $D = (d_1, d_2, \dots, d_k)$ and M such that $0 \leqslant d_1 < d_2 < \cdots < d_k \leqslant N - M$. Then the sequence E_N is considered as a "good" PR sequence if both measures $W(E_N)$ and $C_k(E_N)$ (at least for small k) are "small" in terms of N. This terminology is justified by the fact that for a (truly) random sequence $E_N \in \{-1, +1\}^N$ both $W(E_N)$ and, for fixed k, $C_k(E_N)$ are around $N^{1/2}$ (up to a logarithmic factor) with probability near 1 (which was proved later by Cassaigne, Mauduit and Sárközy [8], and sharpened by Alon, Kohayakawa, Mauduit, Moreira and Rödl [3]). Since that many constructions have been given for binary sequences possessing strong PR properties (in terms of these measures), and further measures of pseudorandomness of binary sequences have been introduced; a survey of all these results has been given by Gyarmati [16].

2. The measures of pseudorandomness of families of binary sequences

In many applications, e.g. in cryptography it is not enough to construct a few "good" PR binary sequences; one usually needs *large families* of them. One of the first constructions of this type was given by Goubin, Mauduit and Sárközy [15]; their construction used the Legendre symbol and polynomials, we will return to it. In the last decades many further large families of "good" PR binary sequences have been constructed. However, in many applications, e.g. in cryptography it is not enough to know that our family of "good" sequences is large; it is more important to know that the family has a rich, complex structure, and the sequences in the family are "independent" and they are "far apart" in a well-defined sense. Thus various measures have been introduced and applied for studying pseudorandomness of families of binary sequences: family complexity, collision, distance minimum, avalanche effect and cross-correlation measure. In this paper my goal is to present a survey of all these measures and their applications.

It should be noted that some of these measures have been deep roots: variants of them have been used since many decades, and they are still intensively studied in certain closely related fields. Some of these fields are (I will also include a reference to an important survey paper or a paper of basic importance in the field): the cross-correlation of order 2 (see the survey papers [13] and [23]); linear complexity for "multisequences" (see [30]); arithmetic crosscorrelations (which started in [14]); merit factor (see [24]); Boolean functions (sequences of length $N = 2^n$ can be identified with Boolean functions; see [7] and [9]). (There are further references in the papers surveyed below.) It would be hopeless to try to overview all these fields and the related results in them. Thus here I will focus on those recent papers in which an attempt has been made to develop a *comprehensive*, constructive and quantitative approach to the pseudorandomness of families of binary sequences.

3. The family complexity

The most important measure of pseudorandomness of families of binary sequences is, perhaps, the *family complexity* which was introduced in 2003 by Ahlswede, Khachatrian, Mauduit and Sárközy [1]:

Definition 1. The *family complexity* or briefly *f-complexity* $\Gamma(\mathcal{F})$ of a family $\mathcal{F}$ of binary sequences $E_N \in \{-1, +1\}^N$ is defined as the greatest integer j so that for any $1 \leqslant i_1 < \cdots < i_j \leqslant N$ and $(\varepsilon_1, \dots, \varepsilon_j) \in \{-1, +1\}^j$ there is at least one sequence $E_N = (e_1, \dots, e_N) \in \mathcal{F}$ which satisfies the "j-specification"

$$e_{i_1} = \varepsilon_1, \dots, e_{i_j} = \varepsilon_j.$$

The f complexity of $\mathcal{F}$ is denoted by $\Gamma(\mathcal{F})$. (If there is no $j \in \mathbb{N}$ with the property above, then we set $\Gamma(\mathcal{F}) = 0$.)

It was explained in [1] in the following way why it is important to know that our family $\mathcal{F}$ of binary sequences constructed by the given PR generator (the so-called "key space") is of high f-complexity:

Assume that $\Gamma(\mathcal{F}) = K$ is a "large" even number, and someone tries to break the code by determining the key sequence (taken from $\mathcal{F}$). Suppose he is able to determine $K/2$ bits of it (at certain positions). Can he use this information? Take any other $K/2$ positions, and consider all the possible ±1 choices at these positions: this gives $2^{K/2}$ possibilities, and by the definition of K, each of these possibilities occurs in at least one sequence of $\mathcal{F}$. Thus there are at least $2^{K/2}$ (exponentially many!) possibilities to extend the known bits into a possible key occurring in the key space, so that the attacker has to check exponentially many possibilities to find the right key! Quoting [1]: "We conclude if we can construct a family $\mathcal{F}$ of high f-complexity and of "good" PR binary sequences, then the cryptosystem based on it (as described above) has good security properties."

Indeed, this consideration was followed in [1] by the construction of such a family (which was a variant of the constructions of Goubin, Mauduit and Sárközy [15] mentioned earlier and of Sárközy and Stewart [32]).

Theorem 1. *Let p be a prime number, $K \in \mathbb{N}$, $L \in \mathbb{N}$ and*

$$(4K)^L < p.$$

Consider all the polynomials $f(x) \in \mathbb{F}_p[x]$ with the properties that

$$0 < \text{degree} f(x) \leqslant K \tag{1}$$

and

$$f(x) \text{ has no multiple zero in } \overline{\mathbb{F}}_p.$$