Set-Codes with Small Intersections and Small Discrepancies

Ryan Gabrys*, H. Dau†, C. J. Colbourn‡, and O. Milenkovic*
*University of Illinois at Urbana-Champaign (UIUC), {gabrys, milenkovic}@uiuc.edu
†Monash University, Melbourne, Australia, hoang.dau@monash.edu
‡Arizona State University colbourn@asu.edu

***Abstract*—We are concerned with the problem of designing large families of subsets over a common labeled ground set that have small pairwise intersections and the property that the maximum discrepancy of the label values within each of the sets is less than or equal to one. Our results, based on transversal designs, factorizations of packings and Latin rectangles, show that by jointly constructing the sets and labeling scheme, one can achieve optimal family sizes for many parameter choices. Probabilistic arguments akin to those used for pseudorandom generators lead to significantly suboptimal results when compared to the proposed combinatorial methods. The design problem considered is motivated by applications in molecular data storage.**

I. Introduction

In his seminal work [2], Beck introduced the notion of the discrepancy of a finite family of subsets over a finite ground set as the smallest integer d for which the elements in the ground set may be labeled by ± 1 so that the sum of labels in each subset is at most d in absolute value. Set discrepancy (bicoloring) theory has since been studied and generalized in a number of different directions [6], [11] and has found applications in pseudorandomness and independent permutation generation [17], ϵ-approximations and geometry [13], bin packing, lattice approximations and graph spectra [16], [19].

The goal of these, and almost all other studies of discrepancies of set families, was to establish bounds on the largest size of families of d-discrepancy sets for a given ground set, or to construct large set families with prescribed discrepancy values. The sets were assumed to have no special structural constraints other than those that ensure desired discrepancy properties. An exception in this context is the work of Colbourn et al. [5] concerning the problem of *bicoloring* Steiner triple systems (STSs) [4]. Steiner triple systems are set systems in which the subsets of interest satisfy additional intersection constraints, ensuring that each pair of distinct elements of the ground set appears in exactly one subset of the system. The key finding is that STSs are inherently impossible to bicolor, as proved in [4].

Recently, the authors proposed a number of coding techniques for molecular storage platforms [8], [10], [14], [18], [23], [24]. In one such paradigm, data is recorded in terms of the locations of nicks (cuts) in naturally occurring DNA strands (see Figure 1 for an illustration). In order to correct readout errors, information is encoded into sets of nicking positions that have small overlaps, i.e., into sets with small intersections. As the DNA-strand is of the form of a double helix, i.e., composed of two strands (a sense and antisense strand), nicks may be introduced on either of the two entities. To prevent disassociation of the strands due to nicking, it is desirable to distribute the nicks equally on both strands. Such a problem setting leads to a constructive set discrepancy problem, in which one desires to construct a large family of subsets of a ground set that have small intersection and small discrepancy d. In this setting, each set gives rise to two codewords in which the labels are complements of each other, as nicks on different strands can be easily distinguished from each other in the sequencing and alignment phase [15].

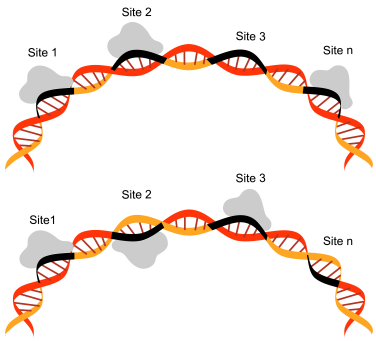

Figure 1: Nicking of native DNA. A nick is a cut in one of the two sugar-phosphate strings, at a designated site. If many closely placed cuts are made only on one strand (top) the DNA may disassociate. To avoid this problem, balancing cuts on both strands is desirable (bottom).

To address the problem, we proceed in two directions. First, we examine existing (near-optimally) sized families of sets with small intersections, such as the Bose-Bush and Babai-Frankl families [1] and Steiner systems [4]. For the former case, we show that one can achieve the smallest possible discrepancy ($d = 0$ for even-sized sets and $d = 1$ for odd-sized sets) in a natural manner, by using the properties of the defining polynomials of the sets. Second, we generalize the results of [5] to show that no Steiner system can have optimal discrepancy. We establish upper bounds on the size of optimal discrepancy intersecting families and then proceed to describe several constructions based on *packings* that have optimal discrepancy values. In the process, we invoke graph-theoretic arguments, and properties of orthogonal arrays. Note that an alternative approach to address the code construction problem for fixed set sizes is to use specialized ternary constant weight codes, but we find the set discrepancy formulation easier

978-1-5386-9291-2/19/$31.00 ©2019 IEEE

to work with and generalize.

The paper is organized as follows. In Section II we precisely formulate the problem and provide simple arguments that show how near-optimal families of sets with small intersection constructed by Bose-Bush and Babai-Frankl may be balanced. Section III is devoted to studying upper bounds on the size of optimal discrepancy families of intersecting sets, while Section IV discusses various set constructions. Due to lack of space, the proofs of all the presented results are omitted, but may be found in the extended manuscript [7].

II. Problem Statement and a Sampling of Results

Let $[v]$ denote the set of integers $\{1, 2, \ldots, v\}$. A family of subsets of $[v]$, $\mathcal{F}_v = \{F_1, \ldots, F_s\}$, $s \geq 2$, is *k-regular* if for all $1 \leq j \leq s$, $|F_j| = k$. Otherwise, the family is *irregular*. The sets in $\mathcal{F}_v$ have *t-bounded intersections* if for all pairs of distinct integers $i, j \in [s]$, $|F_i \cap F_j| \leq t-1$. Naturally, when $\mathcal{F}_v$ is k-regular, we require that $t < k$.

Let $L : [v] \to \{+1, -1\}$ be a labeling of the elements in $[v]$. The *discrepancy* of a set $F_j \in \mathcal{F}_v$ under the labeling L is $D_L(F_j) = \sum_{i \in F_j} L(i)$. For fixed values of v and t, our goal is to find the largest size s of a t-bounded intersection family $\mathcal{F}_v$ for which there exists a labeling L such that $D_L(F_j) \in \{-1, 0, +1\}$ for all $1 \leq j \leq s$. We refer to such a set system as an *extremal balanced family*.

Sets with bounded pairwise intersections have been extensively studied in the past [1], [4]. Well-known examples include Steiner systems, corresponding to k-regular families $\mathcal{F}_v$ with the property that each t-subset of $[v]$ belongs to exactly one member of the family [4], and $\mathcal{C}$-intersecting families of sets described in [1]. In the latter, the cardinalities of the intersections of the sets are restricted to lie in a predetermined set $\mathcal{C}$. Very little is known about discrepancies of intersecting sets and extremal balanced families in particular. The problem at hand is difficult, and it appears hard to construct extremal families for arbitrary parameter values. We therefore mostly focus on some specific choices of the parameter sets.

Some families of t-bounded intersection sets with (near) optimal size have inherently simple labelings that ensure the balancing property. We identify one such family, described in [1], and in what follows describe a simple proof that this family is indeed a balanced family.

Construction [1, Thm. 4.11]. Let q be a prime power and $1 \leq t \leq k \leq q$. Set $v = kq$. Let ξ be a primitive element of the finite field $\mathbb{F}_q$ and set $\mathcal{A} = \{0, 1, \xi, \ldots, \xi^{k-2}\}$, so that $|\mathcal{A}| = k$. For each polynomial $f \in \mathbb{F}_q[x]$, define a set of pairs of elements

$$A_f \triangleq \{(a, f(a)) \colon a \in \mathcal{A}\}.$$

Then $|A_f| = k$. Let

$$\mathcal{C}(k, q) \triangleq \{A_f \colon f \in \mathbb{F}_q[x], \deg(f) \leq t-1\}.$$

Then $\mathcal{C}(k, q)$ is a collection of q^t k-subsets of the set $\mathcal{X} \triangleq \mathcal{A} \times \mathbb{F}_q$ such that every two sets intersect in at most $t-1$ elements, because two distinct polynomials of degree $\leq t-1$ cannot intersect in more than $t-1$ points. When v is large, the Babai-Frankl construction requires $q = v/k$ to be large as well.

The Ray-Chauduri-Wilson Theorem [1] (Theorem 4.10), asserts that the size of any family $\mathcal{F}_v$ of k-regular sets with $k \geq t$ whose pairwise intersection cardinalities lie in some set of cardinality t that satisfies $|\mathcal{F}_v| \leq \binom{v}{t}$. As an example, the set of all t-subsets of $[v]$ forms a $(t-1)$-intersection bounded t-regular family. This result can be strengthened when the set of allowed cardinalities equals $\{0, 1, \ldots, t-1\}$, provided that $v > 2k^2$, for fixed k and t. This is the best known such bound, and it is met by the Babai-Frankl construction described above. It is easy to see that the size of the family is roughly equal to $\frac{v^t}{k^t}$, which has the same growth rate with respect to v as the approximate upper bound of Ray-Chauduri and Wilson $\frac{v^t}{(2k)^t}$, and dominates the other terms provided that v is sufficiently large and k, t are kept constant.

Proposition 1. *There exists a labeling L of points $(a, f(a))$ of the set $\mathcal{X}$ so that every set in $\mathcal{C}(k, q)$ has discrepancy equal to 0 when k is even, and discrepancy equal to ± 1 when k is odd.*

Proof. We prove the statement by constructing suitable labelings for elements of $\mathcal{X}$. When k is even, for every $a \in \mathcal{A}$ and $b \in \mathbb{F}_q$, the pair (a, b) is mapped to -1 for $a \in \{0, 1, \xi, \ldots, \xi^{k/2-2}\} \subset \mathcal{A}$ and 1 for $a \in \{\xi^{k/2-1}, \ldots, \xi^{k-2}\} \subset \mathcal{A}$. Then, every set in $\mathcal{C}(k, q)$ has half of the elements mapped to -1 and half mapped to $+1$. Equivalently, the discrepancy of every set is equal to 0. When k is odd, the pair (a, b) is mapped to -1 for $a \in \{0, 1, \xi, \ldots, \xi^{(k-1)/2-2}\} \subset \mathcal{A}$ and to 1 for $a \in \{\xi^{(k-1)/2-1}, \ldots, \xi^{k-2}\} \subset \mathcal{A}$. The discrepancy of every set is equal to 1 in this case. ∎

A few remarks are in order. First, the labeling presented is solely based on the set $\mathcal{A}$, and the first entry in each pair partitions the set $\mathcal{X}$ into k-groups. Hence, the balancing property is inherited from the partition of $\mathcal{A}$, which suggests a close connection with constructions of *transversal designs.*

A *transversal design* $TD(t, k, v)$[1] consists of

1) A set V of kv elements (called points);
2) A partition of V into sets $\{G_i : i \in [k]\}$, where each G_i contains v points and is called a group;
3) A set $\mathcal{B}$ of k-subsets called blocks for which a) every block and every group intersect in exactly one point ("blocks are transverse to groups"); and b) every t-subset of V either occurs in exactly one block or contains two or more points from a group (but not both).

Because no t-subset of elements can appear in two or more blocks, any two distinct blocks of a $TD(t, k, v)$ intersect in at most $t-1$ elements. It is well known that

- Whenever q is a prime power and $1 \leq t \leq k \leq q+1$, there exists a $TD(t, k, q)$.

[1]A $TD(t, k, v)$ is equivalent to a $OA(t, k, v)$ orthogonal array, which in turn is equivalent to $k-2$ mutually orthogonal Latin squares of order n when $t = 2$ [9]. We refer to all these entities as transversal designs.

- Whenever q is a power of 2 and $1 \leq k \leq q+2$, there exists a $TD(3, k, q)$.
- For any positive integers v and t, both $TD(t, t+1, v)$ and $TD(t, t, v)$ exist.

Whenever a $TD(t, k, v)$ exists, one can assign positive labels to the points in half of the groups and negative labels to the points in the other half of the groups when k is even, or nearly half when k is odd. This automatically leads to a well-defined balanced family of sets, as necessarily every block meets every group in a single point. It is straightforward to see that the Babai-Frankl construction produces a transversal design, as outlined in [20]. However, this construction is not optimal in general, since it may be possible to add k-blocks to the design that intersect the groups in more than one point.

One can add additional k-blocks to the designs as follows. For simplicity, assume that k is even and that $t \geq 3$. Pick one group of the design that lies within the set of positively labeled elements P_+ and one group of the design that lies within the set of negatively labeled elements P_-. There are $\left(\frac{k}{2}\right)^2$ such pairs of groups. By construction, any k-subset with $\frac{k}{2}$ points from the first group and $\frac{k}{2}$ points from the second group intersects each block of the TD in at most two points. Furthermore, each pair of such blocks intersects in at most $\lceil \frac{k}{2} \rceil$ points, so that as long as $t > \max\{\frac{k}{2}, 2\}$, the augmented TD is both balanced and satisfies the intersection constraint[2]. This observation illustrates the fact that extremal balanced families of sets with small intersections cannot be directly derived from TDs.

Henceforth, we focus on investigating the problem of jointly constructing large intersecting families with labelings that ensure that the set discrepancies are contained in $\{0, \pm 1\}$. Since it is known that Steiner triple systems cannot be balanced, i.e., that for any bicoloring of a Steiner triple system there exists one "monochromatic" set (i.e., a set with discrepancy $+3$ or -3) [5], we focus our attention on *packings* instead [4].

A packing $\mathcal{C}(t, k, v)$ with parameters (t, k, v) is a k-regular family of subsets $\mathcal{F}$ of $[v]$ with the property that each t-element subset of $[v]$ appears in *at most* one subset. This automatically ensures that any two distinct $F_i, F_j \in \mathcal{F}_v$ satisfy $|F_i \cap F_j| \leq t-1$. It is customary to refer to the subsets as blocks, and we employ both terms. In the sections to follow, we establish the existence of packings with perfect balancing properties based on explicit constructions that rely on orthogonal arrays and factorizations of graphs [3].

The problem of determining large families of t-bounded intersecting sets has also been independently studied in the theoretical computer science literature, where such sets were considered for generating pseudorandom strings [22]. Most approaches use the probabilistic method. In one such setting [22], the ground set $[v]$ is divided into k disjoint intervals of size $\frac{v}{k}$. A subset $S \subseteq [v]$ is termed "structured" if it contains exactly one element from each interval. A structured set S is generated by picking uniformly at random, with replacement, one element from each interval and adding it to S. A probabilistic argument reveals that there exists a set of $\left(\frac{vt}{k^2}\right)^t$ structured sets, each pair of which intersects in at most t positions. This bound, compared to the Ray-Chauduri-Wilson bound, is smaller by a factor of $\left(\frac{t}{k^2}\right)^t$, but balancing the sets is even easier: points in half of the intervals can be labeled by $+1$ and points in the other half by -1 (or vice versa). If the number of intervals is even, the discrepancy of each set is 0; if the number of intervals is odd, the discrepancy is ± 1.

[2]We can generalize this argument to form new blocks using $s > 2$ groups, half of which are labeled $+1$ and half of which are labeled -1. As long as $t > \max\{\frac{k}{s}, s\}$, the new blocks are valid provided that any two collections of s groups have fewer than $\frac{t}{k/s}$ groups in common. To avoid notational clutter, we assume that $\frac{k}{s}$ is an integer.

III. Upper Bounds

We first derive upper bounds on the size of extremal balanced regular packings (which we refer to as balanced packings for shorthand), and then proceed to establish constructive lower bounds for some given choices of parameters.

For ease of notation, for a given labeling L, let $P_+ = \{i \in [v] : L(i) = +1\}$, $P_- = \{i \in [v] : L(i) = -1\}$, $p_+ = |P_+|$, and $p_- = |P_-| = v - p_+$. Without loss of generality, we assume that $p_+ \geq p_-$. We use $A(t, k, v)$ to denote the largest size of a balanced packing with parameters (t, k, v). The following simple upper bound on $A(t, k, v)$ is based on standard counting arguments.

Lemma 1. *For any labeling L with label classes of size p_+ and p_- such that $p_+ \geq 2\lceil \frac{t+1}{2} \rceil$, and for $t < k$, one has*

$$A(t, k, v) \leq \frac{\binom{p_+}{\lfloor t/2 \rfloor}\binom{p_-}{\lceil t/2 \rceil}}{\binom{\lceil k/2 \rceil}{\lfloor t/2 \rfloor}\binom{\lfloor k/2 \rfloor}{\lceil t/2 \rceil}}.$$

As may be observed from Lemma 1, the maximum size of a regular (t, k, v) packing depends only on the values of p_+ and p_-. The next simple corollary establishes which values of these parameters maximize the upper bound. The obtained bound has the same asymptotic growth as the Ray-Chaudhuri-Wilson bound, $\left(\frac{v}{k}\right)^t$, although it addresses both the intersection *and* discrepancy constraints. It also shows that the construction by Bose-Bush and Frankl-Babai is near extremal in terms of satisfying both joint intersection and balancing conditions, although the construction itself was proposed for addressing intersection constraints only.

Corollary 1. *For $t < k$,*

$$A(t, k, v) \leq \frac{\binom{\lceil v/2 \rceil}{\lfloor t/2 \rfloor}\binom{\lfloor v/2 \rfloor}{\lceil t/2 \rceil}}{\binom{\lceil k/2 \rceil}{\lfloor t/2 \rfloor}\binom{\lfloor k/2 \rfloor}{\lceil t/2 \rceil}}.$$

In what follows, we focus our attention on constructing balanced (t, k, v) packings that meet the bound from Corollary 1 with equality. Given that one can perfectly balance the Babai-Frankl sets, the natural question arises if all or some Steiner systems, in which every t-subset is required to appear in exactly one block, can be perfectly

balanced. Theorem 1 states that the answer to this question is negative: perfectly balanced systems necessarily have cardinalities smaller than that of Steiner systems with the same parameters. This result complements and extends the findings of [5] which pertain to Steiner triple systems only and are considered in the setting of colorability of Steiner systems.

Theorem 1. *For $t < k < v$ where $p_+ > \frac{k+1}{2}$ and $v > 2$,*

$$A(t,k,v) < S(t,k,v).$$

IV. Optimal Code Constructions

Next, we describe extremal balanced intersecting families of sets (i.e., families of sets that meet the bound in Corollary 1) for several parameter choices. In particular, we exhibit extremal balanced set constructions based on Latin rectangles for all values of v and $t = 2$, $k = 3$.

A. Extremal balanced systems with $(t = 2, k = 3, v)$

We consider a simple construction for the parameters $t = 2, k = 3$ is in terms of *factorizations of graphs.* Recall that a *factor* of a graph is a subgraph with the same set of vertices as the graph. If the spanning subgraph is r-regular, it is an *r-factor*. A graph is *r-factorizable* if its edges can be partitioned into disjoint r-factors. Then, a 1-factor of a graph is a *perfect matching*, and a 1-*factorization* is a partition of the graph into matchings. Equivalently, a 1-factorization of a d-regular graph is a proper coloring of the edges with d colors. Suppose that K_{p_+} is a complete graph with vertex set P_+. Let $\Phi = \{\Phi_1, \ldots, \Phi_{p_+-1}\}$ be a 1-factorization of K_{p_+}. Let the triples be of the form $\{i, a_1, a_2\}$, where $i \in [p_-]$ and where the edge $(a_1, a_2) \in \Phi_i$. It is straightforward to see that the resulting system is a partial $(2, 3, v)$ system consisting of triples defined over $[v]$. When $p_- \neq p_+$, we have the following result followed by an example which highlights the main ideas.

Lemma 2. *Suppose that p_+ is even and that $p_- < p_+$. Then, $A(2,3,v) = \frac{p_+ p_-}{2}$.*

Example 1. Let $P_+ = 6$ and $P_- = 3$, and for simplicity, assume that $P_+ = \{1,2,3,4,5,6\}$ and $P_- = \{a,b,c\}$. Then, $\Phi = \{\Phi_1, \ldots, \Phi_5\}$ is a 1-factorization of K_6 where $\Phi_1 = \Big\{\{1,2\},\{3,4\},\{5,6\}\Big\}$, $\Phi_2 = \Big\{\{1,4\},\{2,6\},\{3,5\}\Big\}$, $\Phi_3 = \Big\{\{1,6\},\{2,3\},\{4,5\}\Big\}$, $\Phi_4 = \Big\{\{2,4\},\{1,5\},\{3,6\}\Big\}$, $\Phi_5 = \Big\{\{1,3\},\{2,5\},\{4,6\}\Big\}$. The triples are formed by adding a to each set in Φ_1, adding b to each set in Φ_2, and adding c to each set in Φ_3. This leads to the following triples: $\{a,1,2\},\{a,3,4\},\{a,5,6\}$, $\{b,1,4\},\{b,2,6\},\{b,3,5\}$, $\{c,1,6\},\{c,2,3\},\{c,4,5\}$. Hence, the construction outlined in Lemma 2 achieves the bound $A(2,3,v) = 9$ of Lemma 1.

Next, we turn to the case where $p_+ = p_-$ and p_+ is even. A simple, yet tedious argument reveals that for these parameter choices, one cannot use the factorization approach outlined in Example 1. Hence, we propose a new construction that relies on a special type of Latin rectangles; in our setting, a Latin rectangle is defined as an array of dimension $p_+ \times p_+$ with entries belonging to a set of cardinality $2p_+$ and such that every element appears *at most* once in each row and column of the array. The rows of the array are indexed by elements from $P_+ = \{0, 1, \ldots, p_+ - 1\}$, while the columns are indexed by elements from the same set, but "boxed" $P_- = \Big\{\boxed{0}, \boxed{1}, \ldots, \boxed{p_+ - 1}\Big\}$, so as to distinguish them from the elements in P_+. Our choice of notation is governed by the fact that we will use the values in P_+ and P_- - unboxed or boxed - to describe indices and placements of the elements within the array. An additional requirement on the Latin rectangles is that they do not have *fixed points*, i.e., elements in the array that are equal to the index of their respective row or column. To more precisely describe the fixed point constraint, let $\ell_{i,j}$ denote the element of the Latin rectangle with row index $i \in P_+$ and column index $j \in P_-$. Obviously, triples of the form $\{i, j, \ell_{i,j}\}$ constitute a balanced packing as long as a) there are no fixed points in the array, in which case one would have $\ell_{i,j} = i$ or $\ell_{i,j} = j$ and therefore have the same point repeated twice and b) the triples are distinct. Clearly, only half of the entries of the rectangle may be included in the packing.

The concept of Latin rectangles without fixed points is illustrated by the next example for which $v = 16$ and $p_+ = p_- = 8$.

	$\boxed{0}$	$\boxed{1}$	$\boxed{2}$	$\boxed{3}$	$\boxed{4}$	$\boxed{5}$	$\boxed{6}$	$\boxed{7}$
0	7	5	$\boxed{5}$	$\boxed{4}$	$\boxed{3}$	$\boxed{2}$	3	1
1	6	$\boxed{4}$	$\boxed{3}$	$\boxed{2}$	$\boxed{1}$	4	2	0
2	$\boxed{3}$	$\boxed{2}$	$\boxed{1}$	$\boxed{0}$	5	3	1	7
3	$\boxed{1}$	$\boxed{0}$	$\boxed{7}$	6	4	2	0	$\boxed{2}$
4	$\boxed{7}$	$\boxed{6}$	7	5	3	1	$\boxed{1}$	$\boxed{0}$
5	$\boxed{5}$	0	6	4	2	$\boxed{0}$	$\boxed{7}$	$\boxed{6}$
6	1	7	5	3	$\boxed{7}$	$\boxed{6}$	$\boxed{5}$	$\boxed{4}$
7	0	6	4	$\boxed{6}$	$\boxed{5}$	$\boxed{4}$	$\boxed{3}$	2

.

The above Latin rectangle leads to 32 *distinct* triples. For instance, from the first column of the above Latin rectangle, we can recover the following set of triples: $\{0,7,\boxed{0}\}$, $\{1,6,\boxed{0}\}$, $\{2,\boxed{3},\boxed{0}\}$, $\{3,\boxed{1},\boxed{0}\}$, $\{4,\boxed{7},\boxed{0}\}$, $\{5,\boxed{5},\boxed{0}\}$. It is straightforward to check that the set of unique triples from this Latin rectangle constitutes a balanced $(2,3,16)$ packing of maximum size.

In our extended work, we show how to construct Latin rectangles without fixed points for a wide range of parameter choices, which implies the following result.

Theorem 2. *For any $v \geq 8$, $A(t,k,v) = \left\lfloor \frac{\lfloor \frac{v}{2} \rfloor \lceil \frac{v}{2} \rceil}{2} \right\rfloor$.*

B. Constructions based on transversal designs

Next, we return to the approach outlined at the beginning of our discussion, in which balanced families of sets with small intersections are obtained using transversal designs (akin to the Babai-Frankl method). The next lemma addresses the case $k = 4$ and $t = 3$ by specifying how to add blocks to the design so that the result is optimal.

Lemma 3. *Suppose that there exists a $TD(t,k,v)$ with $v=4m$ and m even. Then,*

$$A(3,4,v)=\frac{\binom{v/2}{2}\binom{v/2}{1}}{2}=\frac{v^2(v/2-1)}{16}.$$

C. Constructions using maximal disjoint Steiner systems

When the sets in $\mathcal{F}$ can be partitioned into classes $\mathcal{F}_1,\ldots,\mathcal{F}_n$ so that $(V,\mathcal{F}_i)$ is a (t',k,v) packing for each $1\leq i\leq n$, we say that $(V,\mathcal{F})$ is t'-*partitionable* with partition classes $\mathcal{F}_1,\ldots,\mathcal{F}_n$. Let $(V_1,\mathcal{F}^1)$ be a (t_1,k_1,v_1) packing that is t_1-partitionable with partition classes $\mathcal{F}_1^1,\ldots,\mathcal{F}_n^1$. Similarly, let $(V_2,\mathcal{F}^2)$ be a (t_2,k_2,v_2) packing that is t_2-partitionable with partition classes $\mathcal{F}_1^2,\ldots,\mathcal{F}_n^2$. Suppose that V_1 and V_2 are disjoint. Form a new packing with blocks $\{B\cup D: B\in\mathcal{F}_i^1, D\in\mathcal{F}_i^2, 1\leq i\leq n\}$. This packing has $v=v_1+v_2$ points. Each block has $k=k_1+k_2$ points. Two distinct blocks can share at most $\max(k_2+t_1-1,k_1+t_2-1)$ points. Hence, the resulting structure is a $(\max(k_2+t_1,k_1+t_2),k_1+k_2,v_1+v_2)$ packing. By setting $V_1=P_+$ and $V_2=P_-$, each block has discrepancy $|k_1-k_2|$. Therefore, we need to choose k_1 and k_2 to have values as close as possible.

1) Balanced sets with parameters $(2,3,v)$: A 1-factorization is a $(2,2,2m)$ packing that is 1-partitionable into $2m-1$ classes. A set of $2m-1$ points, each forming a block of size one, is a $(1,1,2m-1)$ packing that is 0-partitionable into $2m-1$ classes. Hence, using this factorization approach we can obtain a $(2,3,4m-1)$ packing with discrepancy 1 having $m(2m-1)$ blocks, or roughly, $\frac{1}{8}v^2$ blocks. In comparison, a Steiner triple system would have $(4m-1)(4m-2)/6$ blocks, which roughly equals $\frac{8}{3}m^2$ ($\frac{1}{6}v^2$).

2) Balanced sets with parameters $(3,4,v)$: A 1-factorization is a $(2,2,2m)$ packing that is 1-partitionable into $2m-1$ classes. Hence, we have a $(3,4,4m)$ packing with discrepancy 0 and $m^2(2m-1)$ blocks, or roughly, $\frac{1}{32}v^3$ blocks. In comparison, a Steiner quadruple system would have $(4m)(4m-1)(4m-2)/24$ blocks, which roughly equals $\frac{8}{3}m^3$ ($\frac{1}{24}v^3$).

3) Balanced sets with parameters $(4,5,v)$: A 1-factorization is a $(2,2,2m)$ packing that is 1-partitionable into $2m-1$ classes. A *large set of Steiner triple system or maximal disjoint Steiner system* [12], [21] is a $(3,3,v)$ packing that is 2-partitionable into $v-2$ classes, or equivalently, a set of $v-2$ Steiner triple systems that have disjoint block sets. Such a systems exists whenever $(2m+1=)\,v\equiv 1,3 \pmod 6$ (with six exceptions, see [12], [21]). So we obtain a $(4,5,4m+1)$ packing with discrepancy 1 having $(2m-1)m^2(2m+1)/6$ blocks, which is roughly $\frac{1}{384}v^4$ blocks. A Steiner quintuple system would have $(4m+1)(4m)(4m-1)(4m-2)/120$ blocks, which roughly equals $\frac{32}{15}m^4$ ($\frac{1}{120}v^4$).

4) Balanced sets with parameters $(5,6,v)$: A large set of Steiner triple systems is a $(3,3,v)$ system that is 2-partitionable into $v-2$ classes, when $(2m+1=)\,v\equiv 1,3 \pmod 6$ (with six exceptions, see [12], [21]). So we obtain a $(5,6,4m+2)$ packing with discrepancy 0 having roughly $(2m-1)\left(\frac{(2m+1)m}{3}\right)^2\approx\frac{8}{3}m^5$ blocks. A Steiner sixtuple system would have $(4m+1)(4m)(4m-1)(4m-2)/120$ blocks, which roughly equals $\frac{32}{15}m^4$ ($\frac{1}{120}v^4$).

Acknowledgment. The authors gratefully acknowledge discussions with Joao Ribeiro, Imperial College, London. The work was supported by the NSF grants CCF 1526875, 1816913, and 1813729 and the DARPA Molecular Informatics program.

REFERENCES

[1] L. Babai and P. Frankl, *Linear Algebra Methods in Combinatorics with Applications to Geometry and Computer Science*. Department of Computer Science, University of Chicago, New York, 1992.
[2] J. Beck, Balanced two-colorings of finite sets in the square i. *Combinatorica*, 1(4):327–335, 1981.
[3] C. Bujtás and Z. Tuza, Transversal designs and induced decompositions of graphs. *arXiv preprint arXiv:1501.03518*, 2015.
[4] C. J Colbourn and J. H Dinitz, *CRC handbook of combinatorial designs*. CRC press, 2010.
[5] C. J Colbourn, J. H Dinitz, and A. Rosa, Bicoloring Steiner triple systems. *the electronic journal of combinatorics*, 6(1):25, 1999.
[6] B. Doerr and A. Srivastav. Multicolour discrepancies. *Combinatorics, Probability and Computing*, 12(4):365–399, 2003.
[7] R. Gabrys, H.S. Dau, C.J. Colbourn, O. Milenkovic, Set-codes with small intersections and small discrepancies. *available on arXiv*, https://arxiv.org/abs/1901.05559, 2019.
[8] R. Gabrys, H. M. Kiah and O. Milenkovic, "Asymmetric Lee distance codes for DNA-based storage," *IEEE Trans. on Info. Theory*, 63(8), 4982-4995, 2017.
[9] A S. Hedayat, N. James, A. Sloane, and J. Stufken, *Orthogonal arrays: theory and applications*. Springer Science & Business Media, 2012.
[10] H.M. Kiah, G.J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles, *IEEE Trans. on Info. Theory*, 62, no. 6, 3125-3146, 2016.
[11] L. Lovász, J. Spencer, and K. Vesztergombi, Discrepancy of set-systems and matrices. *European Journal of Combinatorics*, 7(2):151–160, 1986.
[12] J-X. Lu, On large sets of disjoint steiner triple systems i,ii, iii, iv. *J. Comb. Theory, Ser. A*, 34(2):140–146, 1983.
[13] J. Matoušek, E. Welzl, and L. Wernisch, Discrepancy and approximations for bounded vc-dimension. *Combinatorica*, 13(4):455–466, 1993.
[14] O. Milenkovic, R. Gabrys, H. M Kiah, and H. Yazdi, Exabytes in a test tube. *IEEE Spectrum*, 55(5):40–45, 2018.
[15] H. Zhao A. Hernandez O. Milenkovic, K. Tabatabei. Nick-based storage in native nucleic acids. *US Patent, UIUC2017-170-02(US) // MBHB 18-1205*, 2018.
[16] T. Rothvoß, Approximating bin packing within o (log opt* log log opt) bins. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 20–29. IEEE, 2013.
[17] M. Saks, A. Srinivasan, S. Zhou, and D. Zuckerman, Low discrepancy sets yield approximate min-wise independent permutation families. *Information Processing Letters*, 73(1-2):29–32, 2000.
[18] O. Milenkovic S.M.T. Yazdi, R. Gabrys, Portable and error-free dna-based data storage. *Scientific Reports*, 7(5011), 2017.
[19] J. Solymosi, Incidences and the spectra of graphs. In *Combinatorial Number Theory and Additive Group Theory*, 299–314. Springer, 2009.
[20] D.R. Stinson, A general construction for group-divisible designs. *Discrete Mathematics*, 33(1):89–94, 1981.
[21] L. Teirlinck, On the maximum number of disjoint triple systems. *Journal of Geometry*, 6(1):93–96, 1975.
[22] L. Trevisan, Extractors and pseudorandom generators. *Journal of the ACM*, 48(4):860–879, 2001.
[23] H Tabatabaei Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, A rewritable, random-access dna-based storage system. *Scientific reports*, 5:14138, 2015.
[24] S. Yazdi, H. M. Kiah, R. Gabrys, and O. Milenkovic, "Mutually uncorrelated primers for DNA-based data storage," *IEEE Trans. on Inform.* Theory, 2018.