PSEUDOSHATTERING PAIRS

NOGA ALON AND VARUN SIVASHANKAR

Abstract. For two vectors $x, y \in [b]^k$, consider the bipartite graph with two copies of $[b]$ in which i on the left is joined to j on the right if $(x_t, y_t) = (i, j)$ for some coordinate t. We study the largest size of a family $C \subseteq [b]^k$ such that, for every two distinct $x, y \in C$, this bipartite graph contains a cycle.

We give a natural construction for such families and conjecture that it is optimal whenever k is large relative to b. We prove an LYM-type upper bound that is asymptotically tight with respect to this construction, and is exact when k is large and divisible by b. We then refine the argument using a circular ordering, obtaining the sharp full-support bound when $k \equiv -1 \pmod{b}$. In the case $b = 3$, we prove the exact general result when $k \equiv -1 \pmod 3$ and k is sufficiently large. The problem is motivated by the Daniely–Shalev-Shwartz dimension and the pseudocube formulation of a higher-alphabet Sauer–Shelah–Perles lemma.

1. Introduction

Let $[m] = \{1, \ldots, m\}$. For a family $\mathcal{F} \subseteq 2^{[n]}$, a set $S \subseteq [n]$ is shattered by $\mathcal{F}$ if

$$\{A \cap S : A \in \mathcal{F}\} = 2^S.$$

Equivalently, if one identifies $\mathcal{F}$ with a binary matrix whose rows are indicator vectors, then a set of columns is shattered if the rows of the corresponding submatrix contain every binary pattern. The Sauer–Shelah lemma [16, 17], in the equivalent formulation due to Pajor [14], states that $\mathcal{F}$ shatters at least $|\mathcal{F}|$ sets.

Recently Hanneke, Meng, Moran, and Shaeiri [8] introduced a b-ary generalization based on pseudocubes. In this paper we focus on the maximum number of *pairs* of columns that are pseudo-shattered. The relevant combinatorial parameter in [8] is the Daniely–Shalev-Shwartz dimension, introduced by Daniely and Shalev-Shwartz [2]. In the classical binary setting, the dual extremal problem of maximizing the number of shattered sets under a fixed row budget was studied by Das and Mészáros [3] and later by Alon, Sivashankar, and Zhu [1]. In particular, the exact maximum for pairs may be deduced by combining a Turán reduction with the theorem of Kleitman and Spencer on pairwise independent set systems [12]. Our problem is the natural b-ary pseudocube analogue of that line of work.

Following [8], a nonempty set $S \subseteq [b]^d$ is called an ℓ-pseudocube if every point of S has at least ℓ neighbors in every coordinate direction. Given a $k \times n$ matrix over $[b]$, a set of d columns is ℓ-pseudo-shattered if the set of row patterns appearing on those columns contains an ℓ-pseudocube. We restrict throughout to the case $d = 2$.

Let us focus on the case $\ell = 1$. Fix a $k \times n$ matrix $M = (m_{t,j})$ over $[b]$. For a pair of columns $u, v \in [n]$, write

$$R_{u,v}(M) = \{(m_{t,u}, m_{t,v}) : t \in [k]\} \subseteq [b]^2.$$

Research supported in part by NSF grants DMS-2154082 and DMS-2349013.

The pair $\{u,v\}$ is pseudo-shattered exactly when $R_{u,v}(M)$ contains a 1-pseudocube. For any $R \subseteq [b]^2$, let $G(R)$ be the bipartite graph with left and right vertex classes both identified with $[b]$, and with an edge ij whenever $(i,j) \in R$. If $S \subseteq R$ is a 1-pseudocube, then every edge of $G(S)$ has another edge with the same left endpoint and another edge with the same right endpoint. Thus every nonisolated vertex of $G(S)$ has degree at least 2, so $G(S)$ contains a cycle, and therefore $G(R)$ contains a cycle. Conversely, the edge set of any cycle in $G(R)$ is a 1-pseudocube. Hence, in dimension two, a pair of columns is 1-pseudo-shattered if and only if its support graph contains a cycle.

Interestingly, this property was studied in the context of phylogeny where pairwise compatibility of multi-state characters is expressed by the acyclicity of a partition-intersection graph; see, for example, Gysel, Lam, and Gusfield [7] and Shutters, Vakati, and Fernández-Baca [18]. At the other extreme, one may require every pair of b-ary columns to realize all ordered pairs of symbols. This stronger qualitative-independence condition was studied by Gargano, Körner, and Vaccaro [6], as well as by Poljak, Pultr, and Rödl [15].

The cycle characterization leads to a clean extremal reformulation. For $x,y \in [b]^k$, write

$$G(x,y) = G(\{(x_t,y_t) : t \in [k]\}).$$

We say that x and y form a *cyclic pair* if $G(x,y)$ contains a cycle. For integers $b,k \geq 2$, let $m_b(k)$ denote the maximum size of a family $C \subseteq [b]^k$ such that every two distinct members of C form a cyclic pair. Let $F_b(n,k)$ denote the maximum number of pseudo-shattered pairs among the columns of a $k \times n$ matrix over $[b]$. The passage from $F_b(n,k)$ to $m_b(k)$ is just Turán's theorem. Given a matrix M, form a graph $H(M)$ on its columns, joining two columns when they are pseudo-shattered. If $H(M)$ contained a clique of size $m_b(k)+1$, then the corresponding column types would give a family in $[b]^k$ of size $m_b(k)+1$ in which every two members form a cyclic pair, contradicting the definition of $m_b(k)$. Hence $H(M)$ is $K_{m_b(k)+1}$-free, and Turán's theorem gives $e(H(M)) \leq t(n, m_b(k))$. Conversely, if we take a largest cyclic clique in $[b]^k$ and repeat its words as evenly as possible as columns, then every two columns from different parts are pseudo-shattered. This realizes the Turán graph, and therefore

$$F_b(n,k) = t(n, m_b(k)),$$

where $t(n,r)$ is the number of edges in the Turán graph $T(n,r)$. Thus the main question is to understand the largest possible cyclic clique, namely $m_b(k)$.

In Section 2 we give a candidate extremal construction and prove that it forms a cyclic clique, giving a lower bound $N_b(k)$ for $m_b(k)$. We believe that this is the exact answer whenever k is sufficiently large relative to b.

The next two sections give evidence for this conjecture from two upper-bound arguments. In Section 3 we prove a path-based permutation-packing LYM-type inequality. This bound applies to arbitrary cliques. It already has the right asymptotic size with respect to the construction in Section 2, and it is exact in the divisible case $k = bq$ under the explicit threshold $(q+1)^q/q! > b$.

In Section 4 we replace the path by a cycle to get a sharper fixed-support bound. For full-support cliques this circular bound matches the good construction when $k \equiv 0, -1 \pmod{b}$. Summing the fixed-support bounds over all possible support sizes gives a general clique bound that improves the path bound eventually, but still has additive lower-support terms. We expect those terms to be artifacts. As evidence, we prove in the specific case

$b = 3$ and $k \equiv -1 \pmod 3$ that the full-support bound already controls an arbitrary clique, giving the exact value of $m_3(k)$ in that residue class for large k.

2. A Lower Bound Construction

Assume $k \geq b$, and write $k = bq + r$, where $0 \leq r < b$. Set

$$s_1 = \cdots = s_{b-r} = q, \qquad s_{b-r+1} = \cdots = s_b = q + 1.$$

Definition 2.1. For a vector $v \in [b]^k$, let

$$A_i(v) = \{t \in [k] : v_t = i\}.$$

We say that v is *good* if

$$|A_i(v)| = s_i \qquad \text{for all } i \in [b],$$

and the first occurrences of the symbols are in increasing order:

$$\min A_1(v) < \min A_2(v) < \cdots < \min A_b(v).$$

Let $N_b(k)$ denote the number of good vectors in $[b]^k$.

Theorem 2.2. *Any two distinct good vectors in $[b]^k$ form a cyclic pair. In particular,*

$$m_b(k) \geq N_b(k).$$

Proof. Let $v, w \in [b]^k$ be distinct good vectors. Write

$$A_i = A_i(v), \qquad B_i = A_i(w).$$

Suppose for contradiction that $G(v, w)$ is a forest.

First remove all isolated edge components from $G(v, w)$. An isolated edge component A_iB_j means that A_i meets only B_j and B_j meets only A_i, which forces $A_i = B_j$ because both $\{A_1, \ldots, A_b\}$ and $\{B_1, \ldots, B_b\}$ are partitions of $[k]$. If every edge component were isolated, then the two partitions would coincide. Since both vectors are good and label their blocks by increasing first occurrence, this would imply $A_i = B_i$ for every i, hence $v = w$, a contradiction. So after removing isolated edge components, a nonempty forest remains; call it H.

If $r = 0$, then every block has size q. Let A_i be a leaf of H, and let B_j be its unique neighbor. Then $A_i \subseteq B_j$, so $|A_i| \leq |B_j|$. Equality is impossible because isolated edges were removed, and strict inequality is impossible because both sizes are q. This contradiction shows that H cannot exist when $r = 0$. Hence we may assume $r > 0$ from now on.

Now let A_i be a leaf of H, with unique neighbor B_j. Since A_i meets no other block on the B-side, we have $A_i \subseteq B_j$. Thus $|A_i| \leq |B_j|$. Equality would again force $A_i = B_j$, giving an isolated edge, so in fact $|A_i| < |B_j|$. Because the only block sizes are q and $q + 1$, we obtain

$$|A_i| = q, \qquad |B_j| = q + 1,$$

and hence

$$i \leq b - r < j.$$

Similarly, if B_ℓ is a leaf of H with unique neighbor A_h, then

$$|B_\ell| = q, \qquad |A_h| = q + 1,$$

so

$$\ell \leq b - r < h.$$

We next claim that H has a leaf on each side. Suppose, for example, that every leaf of H lay on the A-side. Then every B-vertex of H would have degree at least 2. In any tree component of H with a vertices on the A-side and c vertices on the B-side, the number of edges is $a+c-1$. Summing degrees over the B-side therefore gives

$$a + c - 1 \geq 2c,$$

so $a > c$ in each component. Summing over all components shows that H has strictly more vertices on the A-side than on the B-side, which is impossible because we removed isolated edge components in matched pairs. Thus H has a leaf on the B-side as well. By symmetry, it also has a leaf on the A-side.

Choose an A-side leaf A_i with unique neighbor B_j, and a B-side leaf B_ℓ with unique neighbor A_h. From the inequalities above we have

$$i \leq b - r < j, \qquad \ell \leq b - r < h,$$

so in particular

$$i < h, \qquad \ell < j.$$

Let

$$\alpha_t = \min A_t, \qquad \beta_t = \min B_t.$$

Since v and w are good,

$$\alpha_1 < \alpha_2 < \cdots < \alpha_b, \qquad \beta_1 < \beta_2 < \cdots < \beta_b.$$

Because $A_i \subseteq B_j$, we have $\beta_j \leq \alpha_i$. Because $B_\ell \subseteq A_h$, we have $\alpha_h \leq \beta_\ell$. Using $i < h$ and $\ell < j$, we get

$$\alpha_i < \alpha_h, \qquad \beta_\ell < \beta_j.$$

Putting everything together yields

$$\beta_j \leq \alpha_i < \alpha_h \leq \beta_\ell < \beta_j,$$

a contradiction. Therefore $G(v, w)$ cannot be a forest, so it contains a cycle. □

Proposition 2.3. *The number of good vectors is*

$$N_b(k) = \prod_{i=1}^{b-1} \binom{k - \sum_{j<i} s_j - 1}{s_i - 1}.$$

In particular, if $k = bq$, then

$$N_b(k) = \frac{k!}{b!(q!)^b}.$$

Proof. We construct a good vector block by block. The first block A_1 must contain 1, so there are

$$\binom{k-1}{s_1 - 1}$$

choices for A_1. After choosing $A_1, \ldots, A_{i-1}$, let m_i be the smallest unused element. The condition of being good forces $m_i \in A_i$, and we may choose the remaining $s_i - 1$ elements of A_i arbitrarily from the unused elements larger than m_i. At that stage there are

$$k - \sum_{j<i} s_j - 1$$

available choices, so the number of possibilities for A_i is

$$\binom{k-\sum_{j<i} s_j - 1}{s_i - 1}.$$

Multiplying over $i = 1, \ldots, b-1$ gives the formula.

If $k = bq$, then $s_1 = \cdots = s_b = q$, so the product formula gives

$$N_b(k) = \prod_{i=1}^{b-1} \binom{k-(i-1)q-1}{q-1}.$$

A straightforward telescoping simplification yields

$$N_b(k) = \frac{(k-1)!}{(q-1)!^b \prod_{t=1}^{b-1}(k-tq)} = \frac{k!}{b!(q!)^b}.$$

□

3. A Permutation-Packing Upper Bound

In this section we give an upper bound in the spirit of Meshalkin's theorem [13].

For a word $x \in [b]^k$, write

$$k_i(x) = |\{t \in [k] : x_t = i\}|,$$

and

$$b'(x) = |\{i \in [b] : k_i(x) > 0\}|.$$

An *interval partition* of $[k]$ is a partition whose blocks can be ordered $I_1, \ldots, I_s$ so that, for some

$$1 = a_1 < a_2 < \cdots < a_{s+1} = k+1,$$

we have

$$I_j = \{a_j, a_j+1, \ldots, a_{j+1}-1\} \qquad \text{for } j = 1, \ldots, s.$$

Lemma 3.1. *Let P and Q be two interval partitions of $[k]$. Then the bipartite overlap graph with vertex classes P and Q and edges given by nonempty intersection is a forest.*

Proof. Order the blocks as

$$P = \{I_1, \ldots, I_s\}, \qquad Q = \{J_1, \ldots, J_t\},$$

from left to right. We argue by induction on $s+t$. The claim is trivial when $s+t=2$.

Both I_1 and J_1 contain the element 1. If $\max I_1 \le \max J_1$, then $I_1 \subseteq J_1$, so I_1 is a leaf in the overlap graph. Delete the vertex I_1 and remove its points from J_1, deleting J_1 as well if it becomes empty. After translating the remaining ground set, we obtain two interval partitions of a smaller set, so the remaining overlap graph is a forest by induction. Reattaching the leaf I_1 shows that the original overlap graph is also a forest.

The case $\max J_1 \le \max I_1$ is symmetric, with J_1 as the leaf. □

Theorem 3.2 (An LYM-type inequality)**.** *Let $C \subseteq [b]^k$ be a clique. For $x \in C$, let $\Pi(x)$ be the set of permutations $\sigma \in S_k$ such that*

$$x_{\sigma^{-1}(1)} x_{\sigma^{-1}(2)} \cdots x_{\sigma^{-1}(k)}$$

is constant on $b'(x)$ consecutive intervals, one interval for each symbol that appears in x. Then the sets $\Pi(x)$, $x \in C$, are pairwise disjoint. Consequently,

$$\sum_{x\in C} b'(x)! \prod_{i=1}^{b} k_i(x)! \le k!.$$

Proof. Suppose $\sigma \in \Pi(x) \cap \Pi(y)$ for some distinct $x, y \in C$. After applying σ, both words become constant on consecutive intervals. Therefore the partitions of $[k]$ induced by the level sets of x and y become interval partitions. By Lemma 3.1, their overlap graph is a forest. But this overlap graph is exactly $G(x, y)$, contradicting the assumption that x and y form a cyclic pair. Thus $\Pi(x) \cap \Pi(y) = \varnothing$ whenever $x \neq y$ in C.

Since the sets $\Pi(x)$ are pairwise disjoint subsets of S_k, we have

$$\sum_{x\in C} |\Pi(x)| \le k!.$$

For fixed x, choose an ordering of the $b'(x)$ symbols appearing in x, and then order the elements inside each level set. This gives

$$|\Pi(x)| = b'(x)! \prod_{i=1}^{b} k_i(x)!,$$

so the displayed inequality gives the result. □

Theorem 3.3. *Let $C \subseteq [b]^k$ be a clique, and write $k = bq + r$, where $0 \le r < b$. Then the following hold.*

(i) *Every clique satisfies*

$$|C| \le \frac{k!}{\min_{x\in C}\left(b'(x)! \prod_{i=1}^{b} k_i(x)!\right)}.$$

(ii) *If*

$$\frac{(q+1)^q}{q!} > b,$$

then every clique satisfies

$$|C| \le \frac{k!}{b!(q!)^{b-r}((q+1)!)^r}.$$

(iii) *Under the same hypothesis,*

$$m_b(k) \le (1 + O(b/q))N_b(k).$$

In particular, for fixed b,

$$m_b(k) = (1 + o(1))N_b(k)$$

as $q \to \infty$.

(iv) *If the hypothesis in* (ii) *holds and $r = 0$, then*

$$m_b(k) = \frac{k!}{b!(q!)^b}.$$

(v) *Let b^* be a value of the integer y which attains the minimum of the expression $y!\prod_{i=1}^{y} k_i!$ over all integers $1 \le y \le k$ and positive integers k_i satisfying $\sum_{i=1}^{y} k_i = k$. Suppose $k = b^*q + r$ with $0 \le r < b^*$. Then, for every $b \ge 2$,*

$$m_b(k) \le \frac{k!}{b^*!(q!)^{b^*-r}((q+1)!)^r}.$$

If $r = 0$, then for every $b \ge b^*$,

$$m_b(k) = \frac{k!}{b^*!(q!)^{b^*}}.$$

Proof. Put

$$D(x) = b'(x)!\prod_{i=1}^{b} k_i(x)!.$$

Item (i) follows immediately from Theorem 3.2, since

$$|C|\min_{x\in C} D(x) \le \sum_{x\in C} D(x) \le k!.$$

Fix the support size $s = b'(x)$, and let $a_1, \dots, a_s$ be the positive multiplicities of symbols in x. If $a_j \ge a_i + 2$, then replacing (a_i, a_j) by $(a_i + 1, a_j - 1)$ changes the product of factorials by the factor

$$\frac{(a_i+1)!(a_j-1)!}{a_i!a_j!} = \frac{a_i+1}{a_j} < 1.$$

Thus, for fixed s, $D(x)$ is minimized when the s positive multiplicities are as equal as possible.

The full-support balanced value is

$$D_b = b!(q!)^{b-r}((q+1)!)^r = b!(q!)^b(q+1)^r.$$

For $s < b$, the balanced s-support multiplicities are all at least q, so write them as $q + e_1, \dots, q + e_s$, where $e_i \ge 0$ and

$$e_1 + \cdots + e_s = (b-s)q + r.$$

Then the corresponding denominator is at least

$$s!(q!)^s(q+1)^{(b-s)q+r}.$$

Dividing by D_b, we get

$$\frac{s!(q!)^s(q+1)^{(b-s)q+r}}{D_b} = \frac{s!}{b!}\left(\frac{(q+1)^q}{q!}\right)^{b-s} > \frac{s!}{b!}b^{b-s} \ge 1.$$

Hence $D(x) \ge D_b$ for every word x, and item (ii) follows from item (i).

It remains to compare this bound with the construction. Let

$$s_1 = \cdots = s_{b-r} = q, \qquad s_{b-r+1} = \cdots = s_b = q+1, \qquad R_i = s_i + \cdots + s_b.$$

By Proposition 2.3,

$$N_b(k) = \frac{(k-1)!}{s_b!\prod_{i=1}^{b-1}(s_i-1)!\prod_{i=2}^{b-1} R_i}.$$

Therefore

$$\frac{k!}{b!(q!)^{b-r}((q+1)!)^r\, N_b(k)} = \frac{k}{b!}\,\frac{\prod_{i=2}^{b-1} R_i}{\prod_{i=1}^{b-1} s_i}$$

$$\leq \frac{k}{b!}\frac{(b-1)!(q+1)^{b-2}}{q^{b-1}} \leq \frac{k}{bq}(1+O(\frac{b}{q})) \leq 1+O(\frac{b}{q}).$$

Together with the lower bound $m_b(k) \geq N_b(k)$, this proves item (iii). When $r = 0$, Proposition 2.3 gives $N_b(k) = k!/(b!(q!)^b)$, so the upper bound is tight, proving item (iv).

For item (v), the definition of b^* implies that every word $x \in [b]^k$ satisfies

$$b'(x)! \prod_{i=1}^{b} k_i(x)! \geq b^*!(q!)^{b^*-r}((q+1)!)^r.$$

The upper bound follows from item (i). If $r = 0$, the construction using b^* symbols has size $k!/(b^*!(q!)^{b^*})$ and embeds in $[b]^k$ for every $b \geq b^*$. Hence the upper bound is attained. $\square$

Remark 3.4. There are only some values of k for which the minimizer b^* is not unique. In fact, if

$$F_k(y) = y!(q!)^{y-r}((q+1)!)^r, \qquad k = yq + r, \quad 0 \leq r < y,$$

then $F_k(y)$ is unimodal in y, so nonuniqueness forces an adjacent tie. A direct adjacent-ratio calculation shows that the only such values are

$$k \in \{2, 5, 16, 23, 223, 269\}.$$

The divisible case in item (v) can also be recognized explicitly. If $k = Qb^*$, then b^* is a minimizer precisely when

$$\max\left\{1, \left\lceil \frac{Q^Q}{Q!} - 1 \right\rceil\right\} \leq b^* \leq \left\lfloor \frac{(Q+1)^Q}{Q!} \right\rfloor.$$

This follows by comparing $F_k(b^*)$ with its two neighboring values; the unimodality of F_k makes these local inequalities sufficient. We omit the details for brevity.

4. Improving the Upper Bound

The permutation-packing bound in Section 3 applies to arbitrary cliques. In particular, it applies to the special case where every word has full support. If $k = bq + r$ with $0 \leq r < b$, this full-support specialization is

$$P_{b,r}(q) := \frac{k!}{b!(q!)^{b-r}((q+1)!)^r}.$$

The circular argument improves this fixed-support bound by replacing the path with a cycle. This use of cyclic orders is an adaptation of Katona's cycle method, introduced in his proof of the Erdős–Ko–Rado theorem and developed more broadly as a permutation method in extremal set theory [10, 11, 4, 5]. The extra structure here is the set of boundary edges between consecutive level intervals. If two clique members are interval words for the same cyclic order and share a boundary edge, then cutting the circle at that edge turns their level sets into two interval partitions of a line. The interval forest lemma then says that their overlap graph cannot contain a cycle. Thus, for a fixed cyclic order, the boundary sets of clique members are disjoint.

For full-support words, the circular bound is $(bq/k)P_{b,r}(q)$. Thus it is strictly smaller than the path bound whenever $r > 0$. Since the canonical construction itself uses full support, this is the right comparison to make; the circular full-support bound matches the construction when $r = 0$ and when $r = b - 1$. To get back to arbitrary cliques, one can sum the circular bounds over the possible support sizes. This gives a valid general upper bound and eventually improves the general path bound, but it is probably not the truth:

the additive lower-support terms should not be necessary. As evidence, we prove that when $b = 3$ and $k \equiv 2 \pmod 3$, the full-support circular bound already upper-bounds arbitrary cliques. The proof is in Section A.

4.1. **The circular support bound.** An *oriented cyclic order* $\Omega = (v_1, \dots, v_k)$ of $[k]$ is a cyclic listing, considered up to rotation but not reversal. There are $(k-1)!$ such orders. Indices are read modulo k, and the cycle-edges are

$$e_j = \{v_j, v_{j+1}\}, \qquad j \in [k].$$

A nonempty set of the form

$$\{v_a, v_{a+1}, \dots, v_{a+L-1}\}, \qquad L \geq 1,$$

is a *cyclic interval* of Ω; in particular $[k]$ itself is allowed.

For $x \in [b]^k$, write

$$A_i(x) = \{s \in [k] : x_s = i\}.$$

We say that x is an Ω-*interval word* if every nonempty level set $A_i(x)$ is a cyclic interval of Ω. Its boundary is

$$\partial_\Omega x = \{e_j : x_{v_j} \neq x_{v_{j+1}}\}.$$

Thus an Ω-interval word using $b'(x)$ symbols has

$$|\partial_\Omega x| = b'(x).$$

Lemma 4.1. *Fix an oriented cyclic order Ω of $[k]$. Let $C \subseteq [b]^k$ be a clique. If $x, y \in C$ are distinct Ω-interval words, then*

$$\partial_\Omega x \cap \partial_\Omega y = \varnothing.$$

Proof. Suppose that $e \in \partial_\Omega x \cap \partial_\Omega y$. Cut the cycle Ω at the edge e. Since e is a boundary edge for both words, no nonempty level set of either word wraps around the cut. Hence the nonempty level sets of x and y become interval partitions of a linearly ordered set. By Lemma 3.1, their overlap graph is a forest. This overlap graph is exactly $G(x, y)$ after deleting isolated symbol vertices, contradicting the assumption that C is a clique. □

For $2 \leq t \leq b$, let

$$C_t = \{x \in C : b'(x) = t\}$$

be the part of a clique consisting of words that use exactly t symbols.

Theorem 4.2 (Circular support-size bound)**.** *Let $C \subseteq [b]^k$ be a clique, and fix $2 \leq t \leq b$. Write*

$$k = tQ + \rho, \qquad 0 \leq \rho < t.$$

Then

$$|C_t| \leq \frac{Q(k-1)!}{(t-1)!\,(Q!)^{t-\rho}((Q+1)!)^{\rho}}.$$

Proof. Let

$$\mathcal{I}_t = \{(\Omega, x) : \Omega \text{ is an oriented cyclic order of } [k],\ x \in C_t,\ \text{and } x \text{ is an } \Omega\text{-interval word}\}.$$

For fixed Ω, the sets $\partial_\Omega x$ are pairwise disjoint t-subsets of the k cycle-edges, so there are at most $Q = \lfloor k/t \rfloor$ possible words x. Hence

$$|\mathcal{I}_t| \leq Q(k-1)!.$$

On the other hand, if the positive multiplicities of $x \in C_t$ are

$$a_1, \dots, a_t, \qquad a_1 + \dots + a_t = k,$$

then x is an Ω-interval word for exactly

$$(t-1)! \prod_{i=1}^{t} a_i!.$$

Indeed, choose the cyclic order of the t labelled level sets and then order the elements inside each level set. By log-convexity of the factorial function,

$$\prod_{i=1}^{t} a_i! \geq (Q!)^{t-\rho}((Q+1)!)^{\rho},$$

and the double count gives the result. $\square$

4.2. **Full-support cliques.** Write $k = bq + r$, where $0 \leq r < b$ and $q \geq 1$, and let C_b be the full-support part of a clique $C \subseteq [b]^k$. Applying Theorem 4.2 with $t = b$ gives

$$|C_b| \leq U_{b,r}(q) := \frac{q(k-1)!}{(b-1)!(q!)^{b-r}((q+1)!)^r}.$$

The path bound for C_b alone is $P_{b,r}(q)$, so this improves the full-support estimate by the factor bq/k.

We compare this with the size of the canonical construction. Let

$$s_1 = \dots = s_{b-r} = q, \qquad s_{b-r+1} = \dots = s_b = q+1$$

be the canonical multiplicities, and put

$$R_i = s_i + s_{i+1} + \dots + s_b.$$

By Proposition 2.3,

$$N_b(k) = \prod_{i=1}^{b-1} \binom{R_i - 1}{s_i - 1} = \frac{(k-1)!}{s_b! \prod_{i=1}^{b-1}(s_i - 1)! \prod_{i=2}^{b-1} R_i}.$$

Thus

$$\frac{U_{b,r}(q)}{N_b(k)} = \frac{q}{(b-1)!} \frac{\prod_{i=2}^{b-1} R_i}{\prod_{i=1}^{b-1} s_i}.$$

In particular,

$$U_{b,0}(q) = N_b(k) \qquad \text{when } r = 0$$

and

$$U_{b,b-1}(q) = N_b(k) \qquad \text{when } r = b-1,$$

so the circular full-support bound is exact in the divisible residue class and in the top residue class. For the remaining residues, the loss is explicit: if $1 \leq r \leq b-1$, then

$$\frac{U_{b,r}(q)}{N_b(k)} = \frac{r!}{(b-1)!} \prod_{m=r+1}^{b-1} \left(m + \frac{r}{q}\right).$$

This factor is 1 when $r = b-1$, is greater than 1 when $1 \leq r \leq b-2$, and records exactly how far the circular full-support bound remains from the construction in the other residue classes.

4.3. **Arbitrary cliques by summing supports.** Using the fixed-support circular bound for each possible support size gives a bound for unrestricted cliques. For $2 \le t \le b$, write

$$k = tQ_t + \rho_t, \qquad 0 \le \rho_t < t,$$

and set

$$B_t(k) = \frac{Q_t(k-1)!}{(t-1)!\,(Q_t!)^{t-\rho_t}((Q_t+1)!)^{\rho_t}}.$$

Then every clique $C \subseteq [b]^k$ satisfies

$$|C| \le \sum_{t=2}^{b} B_t(k).$$

Indeed, a nontrivial clique contains no word of support size one, and Theorem 4.2 bounds the contribution of each support size $t \ge 2$; the singleton case is trivial.

This summed circular bound is eventually stronger than the path bound in every nondivisible residue class: the full-support term is smaller by the factor bq/k, while all smaller-support terms are exponentially smaller than the full-support term for fixed b. Still, we expect the lower-bound construction to be the exact answer for all congruence classes once k is large enough. The obstacle in this argument is that it treats the support sizes independently and therefore carries additive lower-support terms.

4.4. **The top residue for $b = 3$.** Put $k \equiv 2 \pmod 3$ and $q = (k-2)/3$. Then Theorem 4.2 with $t = 3$ gives

$$|C_3| \le \frac{q(k-1)!}{2q!(q+1)!^2}.$$

This equals

$$N_3(k) = \binom{k-1}{q-1}\binom{2q+1}{q},$$

so (as we have already seen for $k \equiv -1 \pmod b$) the canonical construction is sharp on the full-support part in this residue class.

An arbitrary clique $C \subseteq [3]^k$ decomposes as $C = C_2 \cup C_3$, since a support-one word cannot belong to a nontrivial clique. The summed support bound would therefore add a separate quantity for $|C_2|$ to the full-support estimate above. Since the construction itself has full support, one naturally expects this extra term to be unnecessary. The following theorem proves this in the present case: for sufficiently large q, an arbitrary clique is already bounded by the full-support quantity above.

The proof is a stability refinement of the same cyclic-order count: a support-two word is charged against the deficit it forces in the full-support circular packing. This is analogous in spirit to Hilton–Milner type refinements of Erdős–Ko–Rado and to later applications of Katona's circle method [9, 5].

Theorem 4.3. *Let $k \equiv 2 \pmod 3$, and put $q = (k-2)/3$. If $q \ge 5$, then*

$$m_3(k) = N_3(k).$$

More precisely, every clique $C \subseteq [3]^k$ satisfies

$$|C| \le N_3(k),$$

and if $C_2 \ne \varnothing$, then the inequality is strict.

The proof is given in Section A.

5. Concluding Remarks and Open Problems

The main open question is whether the canonical construction is eventually optimal for every fixed alphabet size. In other words, we expect that for each fixed b,

$$m_b(k) = N_b(k)$$

for all sufficiently large k. The results above prove this asymptotically, prove it exactly when k is large and divisible by b, and prove it exactly for $b = 3$ when $k \equiv -1 \pmod 3$ and k is sufficiently large.

The circular argument suggests a route to the general conjecture. For full-support cliques it gives the sharp bound when $k \equiv 0, -1 \pmod b$, but the unrestricted bound obtained by summing over support sizes still includes additive lower-support terms. We believe these terms are artifacts of treating the support sizes separately. The proof for $b = 3$ and $k \equiv -1 \pmod 3$ shows one possible mechanism: a lower-support word forces enough deficit in the full-support circular packing to pay for itself. A natural next step is to develop such a stability argument for general b.

The other residue classes may require a sharper fixed-support argument. When $1 \le r \le b - 2$, the circular full-support bound remains larger than the construction by an explicit factor, so removing the additive lower-support terms alone would not prove the conjectured exact value. It would be useful to find a refinement of the circular packing, perhaps using additional boundary information, that closes this remaining gap.

One can try to extend the upper bounds proved here for $m_b(k)$ for pairs that are ℓ-pseudo-shattered (for $\ell \ge 1$) in the sense of [8]. In dimension two, this means that the support graph contains a nonempty subgraph of minimum degree at least $\ell + 1$. Let $m_{b,\ell}(k)$ be the maximum size of a family $C \subseteq [b]^k$ with this property for every pair of distinct words; thus $m_{b,1}(k) = m_b(k)$.

It is possible to replace the path and cycle permutation packing arguments used for $\ell = 1$ by embedding in other appropriate graphs. While this gives some nontrivial bounds they appear to be far from optimal. It is worth noting that for $\ell = b - 1$, the condition is qualitative independence, whose sharp asymptotic exponent is known to be $2/b$, as proved by Gargano, Körner, and Vaccaro [6].

Appendix A. The Support-Two Replacement Argument for $b = 3$

Proof of Theorem 4.3. The canonical construction gives the lower bound, so it remains to prove the upper bound. Fix a clique $C \subseteq [3]^k$. If $|C| = 1$, the result is trivial, so assume $|C| > 1$. Then C contains no constant word, because a constant word cannot form a cycle with any other word. Thus

$$C = C_2 \cup C_3.$$

The full-support defect. Let

$$w = 2q!(q+1)!^2, \qquad N = N_3(k) = \frac{q(k-1)!}{2q!(q+1)!^2}.$$

For a full-support word x, let $p(x)$ be the number of oriented cyclic orders Ω for which x is an Ω-interval word. If the three multiplicities of x are a, b, c, then

$$p(x) = 2a!b!c!.$$

Indeed, there are $2 = (3-1)!$ cyclic orders of the three labelled level sets, and then the elements inside each level set may be ordered arbitrarily.

The product $a!b!c!$, subject to $a+b+c=k$ and $a,b,c>0$, is minimized at the balanced multiplicities

$$(q, q+1, q+1).$$

Thus every $x \in C_3$ satisfies

$$p(x) \geq w,$$

with equality exactly for balanced full-support words.

For a cyclic order Ω, put

$$a_3(\Omega) = |\{x \in C_3 : x \text{ is an } \Omega\text{-interval word}\}|\,.$$

By Lemma 4.1, the boundary triples $\partial_\Omega x$, for $x \in C_3$ an Ω-interval word, are pairwise disjoint 3-subsets of the k cycle-edges. Thus

$$a_3(\Omega) \leq q.$$

Summing over all $(k-1)!$ oriented cyclic orders gives

$$\sum_{x \in C_3} p(x) \leq q(k-1)!.$$

Since $wN = q(k-1)!$, this already gives

$$|C_3| \leq N.$$

Define

$$\Delta = N - |C_3| \geq 0, \qquad E = \sum_{x \in C_3} (p(x) - w), \qquad D(\Omega) = q - a_3(\Omega).$$

Since $wN = q(k-1)!$, we have

$$\sum_{\Omega} D(\Omega) = q(k-1)! - \sum_{x \in C_3} p(x) = w\Delta - E.$$

In particular, $E \leq w\Delta$.

Call a full-support word balanced if its three multiplicities are $q, q+1, q+1$ in some order, and call it unbalanced otherwise. Set

$$U(\Omega) = |\{x \in C_3 : x \text{ is unbalanced and } \Omega\text{-interval}\}|\,.$$

Lemma A.1. *We have*

$$\sum_{\Omega} U(\Omega) \leq (q+2)E.$$

Consequently,

$$\sum_{\Omega} \big(D(\Omega) + U(\Omega)\big) \leq (q+2)w\Delta.$$

Proof. Among all unbalanced triples of positive integers with sum k, the product of factorials is minimized at $(q, q, q+2)$: if two parts differ by at least 3, moving one unit from the larger part to the smaller part decreases the product. Therefore every unbalanced $x \in C_3$ satisfies

$$p(x) \geq 2q!q!(q+2)! = w\frac{q+2}{q+1}.$$

Hence

$$p(x) \leq (q+2)(p(x) - w).$$

Summing over the unbalanced full-support words gives

$$\sum_{\Omega} U(\Omega) \le (q+2)E.$$

Combining this with $\sum_{\Omega} D(\Omega) = w\Delta - E$ and $E \le w\Delta$ gives

$$\sum_{\Omega} \big(D(\Omega) + U(\Omega)\big) \le (w\Delta - E) + (q+2)E = w\Delta + (q+1)E \le (q+2)w\Delta.$$

□

A support-two word forces many bad cyclic orders. Now fix $z \in C_2$. Let its two nonempty level sets be A and B, with

$$|A| = a \le |B|\,.$$

Let $p(z)$ be the number of oriented cyclic orders for which z is an interval word. Then

$$p(z) = a!(k-a)!.$$

In particular,

$$p(z) \ge \mu, \qquad \mu = \lfloor k/2 \rfloor! \, \lceil k/2 \rceil!.$$

Call a cyclic order Ω *bad for z* if z is an Ω-interval word and

$$D(\Omega) \ge 1 \qquad \text{or} \qquad U(\Omega) \ge 1.$$

Proposition A.2. *For every $z \in C_2$,*

$$|\{\Omega : \Omega \textit{ is bad for } z\}| \ge \frac{p(z)}{2}.$$

Proof. We split into two cases.

Case 1: $a \le q$. Let Ω be any cyclic order in which z is an interval word. The two boundary edges of z split the circle into two arcs, one corresponding to A and the other to B. The A-arc has only $a-1 \le q-1$ internal cycle-edges.

Let $x \in C_3$ be an Ω-interval word. If x is cyclic with z, then at least two level sets of x must meet both A and B; otherwise the support graph $G(z,x)$, which has only two vertices on the z-side, cannot contain a cycle. Therefore $\partial_\Omega x$ must contain an internal edge of the A-arc.

By Lemma 4.1, these boundary edges are distinct for distinct x. Hence at most $a-1 \le q-1$ full-support words of C can be Ω-interval words. Thus $D(\Omega) \ge 1$, so every interval order of z is bad.

Case 2: $a \ge q+1$. Write

$$a = q+s, \qquad 1 \le s \le q.$$

The upper bound follows from $a \le |B| = 3q+2-a$.

We first record an exact-cover calculation. Index the cycle-edges by

$$0, 1, \ldots, 3q+1 \pmod{3q+2},$$

and suppose the two boundary edges of z are

$$q+s-1 \qquad \text{and} \qquad 3q+1.$$

A balanced full-support interval word has block sizes $q, q+1, q+1$. Therefore its boundary triple has the form

$$T_i = \{i,\ i+q,\ i+2q+1\} \pmod{3q+2}$$

for a unique i.

Suppose q such triples are pairwise disjoint and avoid the two boundary edges of z. Since they use $3q$ edges, they cover every edge except

$$q+s-1, \qquad 3q+1.$$

Let $f_i \in \{0,1\}$ indicate whether T_i is selected. Each edge j belongs exactly to the three triples T_j, T_{j-q}, T_{j-2q-1}, so

$$(*) \qquad f_j + f_{j-q} + f_{j-2q-1} = \begin{cases} 0, & j = q+s-1 \text{ or } j = 3q+1, \\ 1, & \text{otherwise,} \end{cases}$$

with all indices taken modulo $3q+2$. The equation at $3q+1$ gives

$$f_{3q+1} = f_{2q+1} = f_q = 0$$

and the equation at $q+s-1$ gives

$$f_{q+s-1} = f_{s-1} = f_{2q+s} = 0.$$

Starting from these zeros and applying $(*)$ successively along the two chains gives

$$f_0 = f_1 = \cdots = f_{s-2} = 1$$

and

$$f_{q+s} = f_{q+s+1} = \cdots = f_{2q} = 1,$$

while all other f_i's are forced to be 0. Thus the selected indices are exactly

$$(1) \qquad \{0,1,\ldots,s-2\} \cup \{q+s, q+s+1, \ldots, 2q\}.$$

We now pair the interval orders of z. A display

$$[X_1, X_2, \ldots, X_m]$$

means the cyclic order whose consecutive blocks are $X_1, X_2, \ldots, X_m$, with the inherited internal orders.

Subcase $s=1$. Here $|A| = q+1$ and $|B| = 2q+1$. Every interval cyclic order of z can be uniquely written as

$$\Omega = [A_0, a_*, V, U, b_*],$$

where

$$|A_0| = q, \qquad |U| = |V| = q,$$

$a_* \in A$, and $b_* \in B$. The A-block is A_0, a_*, and the B-block is V, U, b_*.

Define a bijection on interval orders of z by

$$\phi(\Omega) = [A_0, b_*, V, U, a_*].$$

The image still has A and B as cyclic intervals: the A-block wraps around the end of the displayed cyclic order.

We claim that Ω and $\phi(\Omega)$ cannot both be not bad. Suppose they were both not bad. Then in each order there are exactly q full-support interval words of C, all balanced. By the exact-cover calculation above, for

$$\Omega = [A_0, a_*, V, U, b_*],$$

the forced family of boundary triples includes the triple whose blocks are

$$P = \{\, U,\ A_0 \cup \{b_*\},\ \{a_*\} \cup V \,\}.$$

Thus P is the level-set partition of some word of C_3.

Similarly, after rotating

$$\phi(\Omega) = [A_0, b_*, V, U, a_*]$$

to the display

$$[a_*, A_0, b_*, V, U],$$

the exact-cover calculation forces the partition

$$Q = \{\, V,\ \{a_*\} \cup U,\ A_0 \cup \{b_*\} \,\}$$

to occur in C_3.

But the overlap graph of P and Q is a forest. Indeed, the common block $A_0 \cup \{b_*\}$ gives an isolated edge component, and the remaining nonzero intersections form the path

$$U - (\{a_*\} \cup U) - (\{a_*\} \cup V) - V.$$

Therefore the corresponding two words are not cyclic, contradicting that C is a clique. Hence at least one of $\Omega, \phi(\Omega)$ is bad. Since ϕ is a bijection, at least half of the interval orders of z are bad.

Subcase $s \geq 2$. Every interval cyclic order of z can be uniquely written as

$$\Omega = [U_A, a_0, A', B', U_B, b_0],$$

where

$$|U_A| = q, \qquad |A'| = s - 1, \qquad |U_B| = q, \qquad |B'| = q + 1 - s.$$

The A-block is U_A, a_0, A', and the B-block is B', U_B, b_0.

Define a bijection by

$$\phi(\Omega) = [U_A, A', B', b_0, U_B, a_0].$$

Again, the image has A and B as cyclic intervals, with the A-block wrapping around the displayed end.

Suppose Ω and $\phi(\Omega)$ are both not bad. For

$$\Omega = [U_A, a_0, A', B', U_B, b_0],$$

the exact-cover calculation forces the partition

$$P = \{\, U_B,\ U_A \cup \{b_0\},\ \{a_0\} \cup A' \cup B' \,\}$$

to occur in C_3.

For $\phi(\Omega)$, rotate the display to

$$[a_0, U_A, A', B', b_0, U_B].$$

The exact-cover calculation then forces

$$Q = \{\, U_A,\ \{a_0\} \cup U_B,\ A' \cup B' \cup \{b_0\} \,\}$$

to occur in C_3.

The overlap graph of P and Q is the path

$$U_B - (\{a_0\} \cup U_B) - (\{a_0\} \cup A' \cup B') - (A' \cup B' \cup \{b_0\}) - (U_A \cup \{b_0\}) - U_A.$$

Hence the two corresponding words are not cyclic, contradiction. Therefore at least one of $\Omega, \phi(\Omega)$ is bad. Since ϕ is a bijection, at least half of the interval orders of z are bad.

This completes the proof of the proposition. $\square$

Counting bad incidences. Let

$$\mathcal{I} = \{(z, \Omega) : z \in C_2,\ \Omega \text{ is bad for } z\}.$$

By Proposition A.2,

$$|\mathcal{I}| \geq |C_2| \frac{\mu}{2}.$$

We now upper-bound $|\mathcal{I}|$. Fix a cyclic order Ω. The support-two words in C_2 that are Ω-interval words have pairwise disjoint boundary pairs, by Lemma 4.1. Therefore there are at most

$$M = \lfloor k/2 \rfloor$$

of them. Moreover, if Ω is bad for some z, then

$$D(\Omega) + U(\Omega) \geq 1.$$

Thus, using Lemma A.1,

$$|\mathcal{I}| \leq M \sum_{\Omega} \big(D(\Omega) + U(\Omega)\big) \leq M(q+2)w\Delta.$$

Consequently

$$|C_2| \leq \frac{2M(q+2)w}{\mu} \Delta.$$

The final numerical estimate. For $q \geq 5$ we have

$$\mu > 2M(q+2)w.$$

Indeed, this is equivalent to

$$R_q := \frac{\lfloor (3q+2)/2 \rfloor! \, \lceil (3q+2)/2 \rceil!}{4 \lfloor (3q+2)/2 \rfloor \, (q+2) q! (q+1)!^2} > 1.$$

Direct computation gives $R_5 = 21/20$ and $R_6 = 9/4$. If

$$n = \lfloor (3q+2)/2 \rfloor, \qquad m = \lceil (3q+2)/2 \rceil,$$

then replacing q by $q+2$ increases both n and m by 3, and cancellation gives

$$\frac{R_{q+2}}{R_q} = \frac{n(n+1)(n+2)(m+1)(m+2)(m+3)}{(q+1)(q+2)^2(q+3)^2(q+4)}.$$

For $q \geq 5$ each numerator factor is larger than the corresponding denominator factor, so $R_{q+2} > R_q$. Thus $R_q > 1$ for all $q \geq 5$.

It follows that $|C_2| < \Delta$ whenever $\Delta > 0$, while if $\Delta = 0$ then $|C_2| = 0$. Hence

$$|C| = |C_3| + |C_2| = N - \Delta + |C_2| \leq N.$$

If $C_2 \neq \varnothing$, then $\Delta > 0$ and $|C_2| < \Delta$, so the inequality is strict. The canonical construction gives the matching lower bound. $\square$

Acknowledgment: Some of the proofs in this paper were found with the help of ChatGPT 5.5 Pro and were subsequently verified by the authors.

References

[1] N. Alon, V. Sivashankar, and D. G. Zhu, *Maximum shattering*, J. Combin. Des. **33** (2025), no. 12, 456–470.

[2] A. Daniely and S. Shalev-Shwartz, *Optimal learners for multiclass problems*, in *Proceedings of the 27th Conference on Learning Theory*, Proceedings of Machine Learning Research, vol. 35, PMLR, 2014, pp. 287–316.

[3] S. Das and T. Mészáros, *Small arrays of maximum coverage*, J. Combin. Des. **26** (2018), no. 11, 487–504.

[4] P. L. Erdős, Z. Füredi, and G. O. H. Katona, *Two-part and k-Sperner families: new proofs using permutations*, SIAM J. Discrete Math. **19** (2005), no. 2, 489–500.

[5] P. Frankl, *Old and new applications of Katona's circle*, European J. Combin. **95** (2021), Article 103339.

[6] L. Gargano, J. Körner, and U. Vaccaro, *Sperner capacities*, Graphs Combin. **9** (1993), 31–46.

[7] R. Gysel, F. Lam, and D. Gusfield, *Constructing perfect phylogenies and proper triangulations for three-state characters*, Algorithms Mol. Biol. **7** (2012), Article 26.

[8] S. Hanneke, Q. Meng, S. Moran, and A. Shaeiri, *An optimal Sauer lemma over k-ary alphabets*, arXiv:2604.12952, 2026.

[9] A. J. W. Hilton and E. C. Milner, *Some intersection theorems for systems of finite sets*, Quart. J. Math. Oxford Ser. (2) **18** (1967), 369–384.

[10] G. O. H. Katona, *A simple proof of the Erdős–Chao Ko–Rado theorem*, J. Combin. Theory Ser. B **13** (1972), no. 2, 183–184.

[11] G. O. H. Katona, *The cycle method and its limits*, in *Numbers, Information and Complexity*, Kluwer Academic Publishers, Boston, 2000, pp. 129–141.

[12] D. J. Kleitman and J. Spencer, *Families of k-independent sets*, Discrete Math. **6** (1973), 255–262.

[13] L. D. Meshalkin, *Eine Verallgemeinerung des Satzes von Sperner über Anzahl der Untermengen einer endlichen Menge*, Teor. Veroyatnost. i Primenen. **8** (1963), 219–220.

[14] A. Pajor, *Sous-espaces ℓ_1^n des espaces de Banach*, Hermann, Paris, 1985.

[15] S. Poljak, A. Pultr, and V. Rödl, *On qualitatively independent partitions and related problems*, Discrete Appl. Math. **6** (1983), 193–205.

[16] N. Sauer, *On the density of families of sets*, J. Combinatorial Theory Ser. A **13** (1972), 145–147.

[17] S. Shelah, *A combinatorial problem; stability and order for models and theories in infinitary languages*, Pacific J. Math. **41** (1972), 247–261.

[18] B. Shutters, S. Vakati, and D. Fernández-Baca, *Incompatible quartets, triplets, and characters*, Algorithms Mol. Biol. **8** (2013), Article 11.

Department of Mathematics, Princeton University, Princeton, New Jersey, USA

Email address: nalon@math.princeton.edu

Department of Mathematics, Princeton University, Princeton, New Jersey, USA

Email address: varunsiva@princeton.edu