Superselectors: Efficient Constructions and Applications

Ferdinando Cicalese and Ugo Vaccaro

Department of Computer Science and Applications "R.M. Capocelli"
University of Salerno, via Ponte don Melillo, 84084 Fisciano, Italy

Abstract. We introduce a new combinatorial structure: *superselectors*. We show that superselectors subsume several important combinatorial structures used in the past few years to solve problems in group testing, compressed sensing, multi-channel conflict resolution and data security. We prove close upper and lower bounds on the size of superselectors and we provide efficient algorithms for their constructions. Albeit our bounds are very general, when they are instantiated on the combinatorial structures that are particular cases of superselectors (e.g., (p, k, n)-selectors [15], (d, ℓ)-list-disjunct matrices [25], $MUT_k(r)$-families [28], $FUT(k, \alpha)$-families [2], etc.) they match the best known bounds in terms of size of the structures (the relevant parameter in the applications). For appropriate values of parameters, our results also provide the first efficient deterministic algorithms for the construction of such structures.

1 Introduction

It is often the case where understanding and solving a problem means discovering the *combinatorics* at the heart of the problem. Equally time and again it happens that the crucial step towards the economical solution of problems arising in different areas hinges on the efficient construction of a *same* combinatorial object. An interesting example is that of superimposed codes [26] (also known as cover-free familes [20], strongly selective families [10], disjunct matrices [16], ...). Superimposed codes represent the main tool for the efficient solution of several problems arising in compressed sensing [11], cryptography and data security [27], computational biology [3], multi-access communication [36], database theory [26], pattern matching [24,34,32], distributed colouring [29], and circuit complexity [4], among the others. Due to their importance, a lot of efforts has been devoted to the design of fast algorithms for the construction of superimposed codes of short length. In this line of research a main result is the paper by Porat and Rotschild [33] who presented a very efficient polynomial time algorithm for that purpose. More recently, Indyk *et al.* [25] showed that optimal nonadaptive group testing procedure (i.e, superimposed codes) can be efficiently constructed and decoded.

In the past few years it has also become apparent that combinatorial structures strictly related to superimposed codes lie at the heart of an even more vast series of problems. As quick examples, the selectors introduced in [9] were instrumental

M. de Berg and U. Meyer (Eds.): ESA 2010, Part I, LNCS 6346, pp. 207–218, 2010.

to obtain fast broadcasting algorithms in radio networks, the (p, k, n)-selectors of [15] were the basic tool for the first two-stage group testing algorithm with an information theoretic optimal number of tests, the (d, ℓ)- disjunct matrices of [25] were a crucial building block for the efficiently decodable non-adaptive group testing procedures mentioned above.

It is the purpose of this paper to introduce *superselectors*, a new combinatorial object that encompasses and unifies all of the combinatorial structures mentioned above (and more). We provide efficient methods for their constructions and apply their properties to the solutions of old and new problems for which constructive solutions have not been shown so far. In particular, superselectors extend at the same time superimposed codes and several different generalizations of theirs proposed in the literature.

When appropriately instantiated, our superselectors asymptotically match the best known constructions of (p, k, n)-selectors [15], (d, ℓ)-list-disjunct matrices [25], monotone encodings and (k, α)-FUT families [31,2], $MUT_k(r)$-families for multiaccess channel [28,1]. In some cases, e.g., for (p, k, n)-selectors and (d, ℓ)-list-disjunct matrices, we also improve on the multiplicative constant in the O notation. We show that optimal size superselectors (and hence all the above structures) can be easily constructed in time polynomial in n, the main dimension of the structure, though exponential in the second parameter p. This might be satisfying in those applications, e.g., computational biology, where $p \ll n$. A major open question is whether it is possible to deterministically obtain optimal size superselectors (or even selectors) in time subexponential in p. However, in cases when p is constant we note that our results provide the first known polynomial construction of *optimal size* (p, k, n)-selectors (and related structures).

It should be also noticed that selectors, and similar combinatorial structures, generally have to be computed *only once*, since they can be successively used in different contexts without the need to recompute them from scratch. Therefore, it seems to make sense (and in absence of better alternatives) to have onerous algorithms that output structures of optimal size, (the crucial parameter that will affect the complexity of algorithms that uses selectors and the like structures in different scenarios) than more efficient construction algorithms that produce structures of suboptimal size. This brings us to another question. Most of the structures mentioned above, and subsumed by our superselectors, can also be obtained via expander graphs, or equivalently, randomness extractors. However, to the best of our knowledge, the best known explicit expander-based constructions give only suboptimal (w.r.t. to the size) selector-like structures. Table 1 summarizes how our results compare to the state of the art. The bounds are reported as they were given in the original papers, thus producing a slight level of difformity. However, if with this choice we might be requiring the reader to put a little bit of effort in the comparisons, we are not risking mistranslations of the bounds from one notation into another. The main aim of the data in the table is to show that the generalization provided by the superselectors in no case implies a loss in terms of optimality of the structure size. In addition, the number of

Table 1. Bounds attained via SUPER-SELECTORS against best known bounds

Structure	Lower Bounds on the Size	Our Upper Bounds on the Size construction time	Old Upper Bounds on the Size, construction time
(p,k,n)-sel.	$\Omega\left(\frac{p^2}{p-k+1}\frac{\log(n/p)}{\log\frac{p}{p-k+1}+O(1)}\right)$ [7,15]	$\frac{2p^2}{p-k+1}\log\frac{n}{p}\,(1+o(1))$ time: $O(n^{p+1}\log n)$	$\frac{p^2}{p-k+1}\,polylog(n)$, time: $poly(n)$ [7] $\frac{ep^2}{p-k+1}\log\frac{n}{p}+\frac{ep(2p-1)}{p-k+1}$, time: $EXP(n)$ [15]
(d,ℓ)-list	$d\log\left(\frac{n}{e(d+\ell-1)}\right)$ if $d<2\ell$ $\frac{d^2}{4\ell\log\frac{ed^2}{4\ell}}\log\frac{n-2\ell-\frac{d}{2}}{ed^2}$ if $d\geq 2\ell$ [15]	$O\left(2d\log\frac{n}{2d}\right),\ d<\ell$ $O\left(\frac{(d+\ell)^2}{\ell}\log\frac{n}{d+\ell}\right),\ d\geq\ell$ time: $O(n^{d+\min\{\ell,d\}+1}\log n)$	$O\left(\frac{2(d+\ell)\log n+\log\binom{d+\ell}{d}}{\frac{\ell}{d+\ell}\left(\frac{d}{d+\ell}\right)^{d/\ell}}\right)$, time: $poly\left(\binom{d+\ell}{d}, n^{d+\ell}, 2^{d+\ell}, \frac{\left(\frac{d}{d+\ell}\right)^{-\frac{d}{\ell}}}{\frac{\ell}{d+\ell}}\right)$ [25]
$MUT_k(p)$	$\Omega\left(\max\{k^2,p\}\log\frac{n}{p}\right)$ [28]	$O((p+k^2)\log\frac{n}{p})$ time: $O(n^{p+1}\log n)$	$O((p+k^2)\log\frac{n}{p})$, non-constructive [1]
(p,α)-FUT	$\frac{p}{(1-\alpha)\log\frac{p}{1-\alpha}}\log n$ [2]	$\frac{p}{(1-\alpha)}\log\frac{n}{p}$ time: $O(n^{p+1}\log n)$	$\frac{p}{(1-\alpha)}\log\frac{n}{p}$, non-constructive [2]
p-cover free	$\left(\frac{p^2}{2\log p}\log\frac{n}{p}\right)(1+o(1))$ [17]	$\frac{ep^2}{\log e}\log\frac{n}{p}(1+o(1))$ time: $O(n^{p+1}\log n)$	$p(p+1)\log e\log\frac{n}{p}$, non-constructive [18,6] $\Theta(p^2\log n)$, time: $\Theta(pn\log n)$ [33].
$(p,\mathbf{v},n)$-sel.	$\max_j \frac{j^2}{j-v_j+1}\frac{\log(n/j)}{\log\frac{j}{j-v_j+1}+O(1)}$	$\max_{j=1,\ldots,p}\{\min\{\frac{j^2}{\log_2 e},\frac{3pej}{j-v_j+1}\}\log\frac{n}{j}\}$	

applications of superselectors we shall present in Section 3 seems to suggest that they represent a basic structure, likely to be useful in many contexts.

2 The $(p, \mathbf{v}, n)$-SUPER-SELECTOR

Given two vectors $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$, we denote with $\mathbf{x} \oplus \mathbf{y}$ the Boolean sum of $\mathbf{x}$ and $\mathbf{y}$, i.e., their componentwise OR. Given an $m \times n$ binary matrix M and an n-bit vector $\mathbf{x}$, we denote by $M \odot \mathbf{x}$ the m-bit vector obtained by performing the Boolean sum of the columns of M corresponding to the positions of the 1's in $\mathbf{x}$. That is, if $\mathbf{x}$ has a 1 in positions, say $3, 7, 11, \ldots$, then $M \odot \mathbf{x}$ is obtained by performing the $\oplus$ of the 3rd, 7th, 11th, ..., column of M. Given a set $S \subseteq [n]$, we use $M(S)$ to denote the submatrix induced by the columns with index in S. Also we use $\mathbf{a}_S$ to indicate the Boolean sum of the columns of $M(S)$. Given two n-bit vector $\mathbf{x}, \mathbf{y}$ we say that $\mathbf{x}$ is *covered* by $\mathbf{y}$ if $x_i \leq y_i$, for each $i = 1, \ldots, n$. Note that if $\mathbf{x}$ is not covered by $\mathbf{y}$ then it means that $\mathbf{x}$ has a 1 in a position in which $\mathbf{y}$ has a 0.

We first recall the definition of (p,k,n)-selector, as given in [15]. A (p,k,n)-selector is an $m \times n$ binary matrix such that for any subset S of $p \leq n$ columns, the submatrix $M(S)$ induced by S contains at least $k \leq p$ rows of the identity matrix I_p. The parameter m is the *size* of the selector.

Definition 1. *Fix integers n, p, with $p \leq n$ and an integer vector, $\mathbf{v} = (v_1, \ldots, v_p)$, such that $v_i \leq i$, for each $i = 1, \ldots, p$. We say that an $m \times n$ binary matrix M is a $(p, \mathbf{v}, n)$-*SUPER-SELECTOR *if M is a (i, v_i, n)-selector for each $i = 1, \ldots, p$. We call m the size of the* SUPER-SELECTOR.

Our main result on SUPER-SELECTORS is summarized in the following theorem, whose proof will be given in Sections 4.

Theorem 1. *A* $(p, \mathbf{v}, n)$-SUPER-SELECTOR *of size*

$$m = O(\max_{j=1,\ldots,p} k_j \log(n/j)), \qquad \text{where } k_j = \min\left\{\frac{3pej}{(j - v_j + 1)}, \frac{ej^2}{\log_2 e}\right\}$$

can be constructed in time polynomial in n *and exponential in* k.

The "identification" capability of a SUPER-SELECTOR are as follows.

Lemma 1. *Let* M *be a* $(p, \mathbf{v}, n)$-SUPER-SELECTOR, $\mathbf{v} = (v_1, \ldots, v_p)$. *Let* S *be any set of* $x < v_p$ *columns of* M. *Let* $\mathbf{a}_S$ *denote the Boolean sum of the columns in* S. *Then, from* $\mathbf{a}_S$ *it is possible to identify at least* v_{x+y} *of the columns in* S, *where* y *is the number of columns of* M *which are not in* S *but are covered by* $\mathbf{a}_S$. *Moreover,* $y < \min\{j \mid x < v_j\} - x$.

Proof. Let $T = \{\mathbf{b} \mid \mathbf{b} \notin S \text{ and } \mathbf{b} \oplus \mathbf{a}_S = \mathbf{a}_S\}$, i.e., T is the set of columns not in S but covered by $\mathbf{a}_S$. Then, $y = |T|$. We first prove the last statement.

Claim. $y < \min\{j \mid v_j > x\} - x$. Let j^* be a value of j achieving the minimum. The claim is a consequence of M being a (j^*, v_{j^*}, n)-selector. To see this, assume, by contradiction, that $|T| \geq j^* - x$. Let $T' \subseteq T$ and $|T' \cup S| = j^*$. Then, there are at least $v_{j^*} > |S|$ columns in $T' \cup S$ with a 1 in a row where all the other columns have a 0. Thus, there is at least one column of T' which has a 1 where all the column of S have a 0. This contradicts the fact that all the columns of T (and hence of T') are covered by $\mathbf{a}_S$.

Since $x+y < j^* \leq p$, and M is an $(x+y, v_{x+y}, n)$-selector, among the columns of $S \cup T$ there are at least v_{x+y} which have a 1 where all the others have a 0. Let W be such set of columns. By an argument analogous to the one used in the claim we have that $W \subseteq S$ and we can identify them. □

Remark 1. Notice that if $v_i > v_{i-1}$, for each $i = 2, \ldots, p$, then we have a situation that, at a first look, might appear surprising: the larger is the number of spurious elements, i.e., columns not in S but covered by $\mathbf{a}_S$, the more information we get on S, i.e., the more are the columns of S that are identified.

Remark 2. The same argument used in the proof above shows that Lemma 1 also holds when $\mathbf{a}_S$ is the component-wise arithmetic sum of the columns in S.

3 Applications of the SUPER-SELECTORS

Approximate Group Testing. In classical non-adaptive group testing [16], we want to identify a subset $P \subseteq [n]$, with $|P| \leq p$, by using the minimum possible set of tests $T_1, \ldots, T_m$, where for each $i = 1, \ldots, m$, we have $T_i \subseteq [n]$. The outcome of test T_i is a bit which is 1 iff $T_i \cap P \neq \emptyset$. If we require that the whole P is identified exactly, and non-adaptively, then it is known that $\Omega(\frac{p^2}{\log p} \log \frac{n}{p})$ tests are necessary [16].

Cheraghchi [8], in the context of error-resilient group testing, Gilbert *et al.* [22], in the context of sparse signal recovery, and Alon and Hod [2] considered

the case when one is interested in identifying some approximate version of P. It turns out [8] that at least $p \log \frac{n}{p} - p - e_0 - O(e_1 \log \frac{n-p-e_0}{e_1})$ tests are necessary if one allows the identification algorithm to report a set P', such that $|P' \setminus P| \leq e_0$ and $|P \setminus P'| \leq e_1$. In other words, the algorithm can report up to e_0 false positives and up to e_1 false negatives.

Let M be an appropriate $(p + e_0, \mathbf{v}, n)$-SUPER-SELECTOR, with the components of vector $\mathbf{v}$ defined by $v_i = i - \min\{e_0, e_1\} + 1$. We can use M to attain approximate identification in the above sense. Proceeding in a standard way, map $[n]$ to the indices of the columns of the super-selector and interpret the rows of the super-selector as the indicator vectors of the tests. Now the vector of the outcomes of the tests is the Boolean sum $\mathbf{a}_P$ of those columns whose index is in P. Let P' be the set of the indices of the columns covered by $\mathbf{a}_P$. We have $P \subseteq P'$ and by Lemma 1 also $|P'| \leq |P| + e_0$. Moreover, from Lemma 1 we also know that a set of positives $P'' \subseteq P$ can be exactly identified, with $|P''| \geq |P| - e_1$. Therefore, any set P^* with $P'' \subseteq P^* \subseteq P'$ satisfies the bounds on the false positives and false negatives.

Note that, for the interesting case of $e_0, e_1 = \Theta(p)$, the above group testing strategy is best possible since it uses $O(p \log \frac{n}{p})$ tests which matches the lower bound of [8]. Cheraghchi [8] considers the case when some tests migh be erroneous and only focuses on the case of zero false negatives. Alon and Hod [2] consider the case of zero false positives and obtain $O(p \log(n/p))$ tests procedures, which are in fact optimal for this case. Gilbert *et al.* [22] allow both false positives and false negatives but their procedures uses $O(p \log^2 n)$ tests. Moreover, our implementation guarantees the exact identification of at least $p' - \min\{e_0, e_1\} + 1$ positives, where $p' \leq p$ is the actual number of positive elements.

Additive Group Testing. We now consider exact group testing with *additive* tests. In this variant, the outcome of testing a subset T_i is the number of positives contained in T_i, i.e., the integer $|T_i \cap P|$.

It is known that $\Omega(\frac{p}{\log p} \log \frac{n}{p})$ tests are necessary if we want to exactly identify P using additive tests (see, e.g., [23] and references therein).

Proceeding analogously to the case of Approximate Group Testing, we can reformulate the additive group testing problem as follows: given positive integers n and $p < n$, minimize the number m of rows of an $m \times n$ 0-1 matrix M such that any set P of up to p columns of M can be identified from their sum[1] $\mathbf{a}_P$.

Let M be an appropriate $(2p, \mathbf{v}, n)$-SUPER-SELECTOR, with the components of vector $\mathbf{v}$ defined by $v_i = i$, for $i = 1, \ldots, \sqrt{p}$ and $v_i = \lceil \frac{i}{2} \rceil + 1$, for $\sqrt{p} < i \leq 2p$. We show that M provides a non-adaptive strategy for additive group testing with $O(p \log(n/p))$ tests.

If $|P| < \sqrt{p}$, using the fact that $v_{|P|+1} = |P| + 1$, Lemma 1 and Remark 2 imply that from $\mathbf{a}_S$ we can identify the whole set P.

If, otherwise, $|P| \geq \sqrt{p}$, by using the fact that $v_{2|P|} > |P|$, by Lemma 1 and Remark 2, from $\mathbf{a}_P$ we can uniquely identify a subset R of P, such that $|R| \geq p/2$ and confine the elements of $P_1 = P \setminus R$ into a set S_1 such that $|S_1| \leq p$. In

[1] Here sum is meant in the arithmetic way, i.e., $\mathbf{z} = \mathbf{x} + \mathbf{y}$ iff $z_i = x_i + y_i$, for each i.

particular $S_1 \cup R$ is the set of all columns of M which are component-wise not larger than $\mathbf{a}_P$.

Now, let $\mathbf{a}_{P_1} = \mathbf{a}_P - \sum_{i \in R} \mathbf{c}_i$, where $\mathbf{c}_i$ denotes the ith column of M and the additions and subtractions among vectors are meant component-wise. Clearly, $\mathbf{a}_{P_1}$ is the sum of P_1, i.e., the columns that are still to be identified. Note also that $\mathbf{a}_{P_1}$ can be computed from $\mathbf{a}_P$ and the set R of identified columns *without* any additional test.

We have now a smaller instance of the same problem from which we started, namely identifying the columns of P_1, among the ones in $M(S_1 \setminus R)$, from their sum $\mathbf{a}_{P_1}$. Also notice that Lemma 1 still applies to the columns of $M(S_1 \setminus R)$. Therefore, repeatedly using the above argument we can eventually identify the whole set P. Again, no additional tests are required since we reinterpret, so to speak, the tests outcomes in light of new acquired knowledge.

Finally, by Theorem 1 a SUPER-SELECTOR M of size $O(p \log \frac{n}{p})$ can be constructed in time $O(n^p)$, which gives the desired result. We hasten to remark that in [23] Grebinsky and Kucherov prove the existence of matrices M with an optimal $O(\frac{p}{\log p} \log \frac{n}{p})$ number of rows for the Additive Group Testing described above. However, it's not clear whether their probabilistic construction can be derandomized, and at which cost. We thought worthwhile to mention that our combinatorial tool gives, for free, a solution to the Additive Group Testing problem using number of tests that differ from the optimal one for only a factor of $\log p$.

Monotone Encodings. Moran *et al.* posed the problem of efficiently constructing (n, k)-monotone encodings of size r, (denoted by $ME(n, k, r)$), i.e., monotone injective functions mapping subsets of $[n]$, of size up to k, into $2^{[r]}$ [31]. Monotone encodings are relevant to the study of tamper-proof data structures and arise also in the design of broadcast schemes in certain communication networks A simple counting argument shows that $ME(n, k, r)$ can only exist for $r = \Omega(k \log n/k)$. We can use our SUPER-SELECTOR for obtaining $ME(n, k, O(k \log n/k))$ in the following way. Let $M^{[t]}$ denote the $(t, \mathbf{v}, n)$-SUPER-SELECTOR defined by the vector $\mathbf{v}$ whose ith component is $v_i = \lfloor i/2 \rfloor + 1$ for each $i = 1, \ldots, t$. By Lemma 1, we have that for any $S \subseteq [t/2]$, from $\mathbf{a}_S$ we can identify at least $|S|/2$ of the columns in $M^{[t]}(S)$. Let S^{yes} (resp. S^{no} be the subset of these columns which we can (resp. cannot) identify from $\mathbf{a}_S$.

We can obtain our mapping in the following way. Given $S_0 \in \binom{[n]}{\leq k}$, we map it to the concatenation of the vectors $\mathbf{a}_0 \mathbf{a}_1, \ldots, ..., \mathbf{a}_{\log k}$, where $\mathbf{a}_i$ is the Boolean sum of the columns of $M^{[k/2^{i-1}]}(S_i)$, with $S_i = S^{no}_{i-1}$.

The mapping is of size $\sum_{j=0}^{\log k} \frac{2k}{2^j} \log \frac{n2^j}{2k} = O(k \log n/k)$, therefore of optimal size. Moreover, by observing that for each $S \subseteq T$ we have $\mathbf{a}_S \leq \mathbf{a}_T$ and $S^{no} \subseteq T^{no}$, we also have that the mapping is monotone. By our Theorem 1 such mapping can be deterministically computed in $O(n^k)$-time.

Alon and Hod [2] defined (k, α)-FUT families in order to obtain $ME(n, k, O(k \log \frac{n}{k}))$ in a way analogous to the one we depicted above, i.e, by

chaining $(\frac{k}{2^t}, \frac{1}{2})$-FUT families[2] of cardinality n for $t = 0, 1, \ldots, \log k$. However, for optimal, i.e., $O(k \log n/k)$-size monotone encodings no explicit deterministic construction has been provided so far [2,31].

Selector-based data compression. Let M be a $(p+1, 2p, n)$-selector of size $m = O(p \log(n/p))$. Let $\mathbf{x}$ be a binary vector with $||\mathbf{x}||_0 \leq p$. Define the encoding of $\mathbf{x}$ as the vector $\mathbf{y}$ equal to the componentwise OR of columns of M corresponding to the positions of the 1's in $\mathbf{x}$. Let $x_{i_1}, \ldots, x_{i_d}$, $d \leq p$, be all the components of $\mathbf{x}$ such that $x_{i_1} = \ldots = x_{i_d} = 1$. By Lemma 1, there exist at most t other columns $m_{j_1}, \ldots, m_{j_t}$ of matrix M, $t \leq p$, such that $\mathbf{y} = m_{j_1} \vee \ldots \vee m_{j_t} \vee m_{i_1} \vee \ldots \vee m_{i_d}$.

Now, think of an "encoder" that works as follows: for a given vector $\mathbf{x}$ it first computes its encoding $\mathbf{y}$, then it computes $A = \{i_1, \ldots, i_d\}, B = \{j_1, \ldots, j_t\}$, and subsequently it computes an ordered list L from $A \cup B$. Finally, the encoder computes a binary vector $\mathbf{z}$ of length $2p$ such that $z_k = 1$ if and only if the k-th element of the ordered list L is an element of A. The encoding of $\mathbf{x}$ is now the concatenated binary vector $\mathbf{yz}$ of length $O(p \log(n/p)) + 2p = O(p \log(n/p))$. One can see that $\mathbf{x}$ can be (efficiently) recovered from $\mathbf{yz}$ and that the length of the encoding $\mathbf{yz}$ of $\mathbf{x}$ is information theoretically optimal.

An extension of the above reasoning can be carried out also to a scenario where $\mathbf{x}$ is generated by a probabilistic source, provided that $Pr\{||\mathbf{x}||_0 > p\}$ goes to zero as the length n of $\mathbf{x}$ grows.

The above encoding procedure has some features which might be of some interest in the area of data compression. Specifically, it does not require construction of code dictionary, nor it is based on statistical analysis of the sequences to be compressed. Moreover, the encoding/decoding procedure only involves simple operations on Boolean vectors (OR's of them and checks for containments), which leads to fast implementation. Furthermore, the above procedure provides a faster alternative for optimal size enumerative encoding of low-weight binary sequences. [12,35]. In particular, for binary vectors of Hamming weight at most d, our encoding/decoding procedures require time $O(nd \log(n/d))$, whereas the procedures given in [35] require time $O(n \log^2 n \log \log n)$ for the encoding, and time $O(n \log^3 n \log \log n)$ for the decoding.

Tracing many users (or finding many positives). In [28] the authors introduced k-out-of-r Multi User Tracing families, aka $MUT_k(r)$. A family $\mathcal{F}$ of n many subsets of $[m]$ is $MUT_k(r)$ if given the union of $\ell \leq p$ of the sets in $\mathcal{F}$, one is able to identify at least k of them, or all if $\ell < k$. Such definition is motivated by applications in multiple access channel communication and DNA computing (see [28] and references quoted therein).

In [1] it was proved that $MUT_k(r)$ families exist for $m = O((r + k^2) \log \frac{n}{r})$, determining the maximum possible *rate* $\frac{\log n}{m}$ for all $k \leq \sqrt{r}$ up to a constant factor. Somehow surprisingly, in all this range the rate is $\Theta(\frac{1}{r})$, independently of k. However, no constructive proof of such "optimal" rate families has been provided so far.

[2] In fact, via SUPER-SELECTORS, we can provide constructions of optimal size (k, α)-FUT families, for any $1/2 < \alpha < 1 - \frac{1}{k}$.

We can use our SUPER-SELECTORS to match such result: Let M be a $(2r, \mathbf{v}, n)$-SUPER-SELECTOR where the vector $\mathbf{v} = (v_1, \ldots, v_{2r})$ is defined by: $v_i = i$ for $i = 1, \ldots, k$; $v_i = k$, for $i = k+1, \ldots, 2r-1$, and $v_{2r} = r+1$.

First, we notice that M is a (k, k, n)-selector, i.e., a $(k-1)$-superimposed code, hence every union of up to $k-1$ columns is unique. Moreover, for any $k \leq \ell \leq r$, by Lemma 1 we have that at least k columns out of ℓ can be identified by their Boolean sum. These two properties show that the sets whose indicator vectors coincide with the columns of M, form an $MUT_k(r)$ family. Therefore, Theorem 1 applied to M provides the best known bound on the size of $MUT_k(r)$ families, i.e., the $O(\max\{r, k^2\} \log n/r)$ of [1]. Our main theorem also explicitly shows that the result of [1] can be attained by a constructive $O(n^k)$ strategy.

The (d, ℓ)-list disjunct matrices. Indyk *et al.* [25] studied (d, ℓ)-list disjunct matrix which are $m \times n$ binary matrix such that the following holds: for any disjoint subsets S, T of columns, such that $|S| \leq d$ and $|T| \geq \ell$, there exists a row where there is a 1 among the columns in T, while all the columns in S have a 0. Such structure was also considered in [14,15,19,8].

One can easily verify that a $(d+\ell, d+1, n)$-selector is also a (d, ℓ)-list disjunct matrix. As a consequence, our Lemma 3 (below) provides improved bounds on construction of (d, ℓ)-list disjunct matrices[3] compared to the ones given in [25].

For any $d \geq \ell$, by using $(d+\ell, d+1, n)$-selector, we obtain (d, ℓ)-list disjunct matrices of size $O(\frac{(d+\ell)^2}{\ell} \log \frac{n}{\ell})$ for any constant d and ℓ. This improves on [25], particularly for d large compared to ℓ. Also for $\ell = \Theta(d)$ and particularly for (d, d)-list disjunct matrices our bound compares favorably with the $O((d \log n)^{1+o(1)})$ size bound given in [25] and the $O(d^{1+o(1)} \log n))$ size bound given in [8]. Alternatively, for $d < \ell$ one can see that a $(2d, d+1, n)$ selector is also a (d, ℓ)-list disjunct matrix. Such a selector can be constructed of size $O(d \log n/d)$, in time $n^{2d+o(1)}$.

We remark that the above results on the size of (d, ℓ)-list disjuct matrices via selectors, are tight with respect to the lower bounds provided in [15, Theorem 2], as reported in Table 1.

4 Bounds on the Size of a $(p, \mathbf{v}, n)$-SUPER-SELECTOR

In this section we prove the bound on the size of a $(p, \mathbf{v}, n)$-SUPER-SELECTOR as announced in Theorem 1. First we present an immediate lower bound following from the ones of [7,15] on the size of (p, k, n)-selectors.

Theorem 2. *The size of a $(p, \mathbf{v}, n)$-SUPER-SELECTOR has to be*

$$\Omega\left(\max_{j=1,\ldots,p} \frac{j^2}{j - v_j + 1} \frac{\log(n/j)}{\log\left(j/(j - v_j + 1)\right) + O(1)}\right).$$

[3] Analogous bounds, in terms of size, are derivable from [15] via (p, k, n)-selectors. However, their construction time is exponential in n.

For the upper bound, we first give a proof based on the probabilistic method and then derandomize it. We need the following two lemmas.

Lemma 2. *There exists a* $(p, \mathbf{v}, n)$-SUPER-SELECTOR *of size*

$$m = O\left(\max_{j=1,\ldots,p} \frac{3pej}{(j - v_j + 1)} \log(n/j)\right).$$

Proof. Generate the $m \times n$ binary matrix M by choosing each entry randomly and independently, with $Pr(M[i,j] = 0) = (p-1)/p = x$. Fix an integer $j \leq p$. Fix $S \in \binom{[n]}{j}$. For any subset R of $j - v_j + 1$ rows of I_j let $E_{R,S}$ be the event that the submatrix $M(S)$ does not contain *any* of the $(j - v_j + 1)$ rows of R. We have

$$Pr(E_{R,S}) = \left(1 - (j - v_j + 1)x^{j-1}(1-x)\right)^m \tag{1}$$

Let $R_1, \ldots, R_t$, $t = \binom{j}{j - v_j + 1}$ be all possible subsets of exactly $j - v_j + 1$ rows of the matrix I_j, and let N_S be the event that, for some index $i \in \{1, \ldots, t\}$, the sub-matrix $M(S)$ does not contain *any* of the rows of the subset R_i. By the union bound we have

$$Pr(N_S) = Pr\left(\bigvee_{i=1}^{t} E_{R_i,S}\right) \leq \binom{j}{j - v_j + 1} \left(1 - (j - v_j + 1)x^{j-1}(1-x)\right)^m \tag{2}$$

One can see that N_S coincides with the the event that the sub-matrix $M(S)$ contains strictly less than v_j rows of I_j. To see this, it is enough to observe that if $M(S)$ contains less than v_j rows of I_j it means that there is some i such that $M(S)$ does not contain any of the rows in R_i.

Let Y_M denote the event that the matrix M is a $(p, \mathbf{v}, n)$-SUPER-SELECTOR. We can use again the union bound to estimate the probability of the negated event $\overline{Y_M}$. If M is not a $(p, \mathbf{v}, n)$-SUPER-SELECTOR then there exists an integer $j \in [p]$ such that for some $S \in \binom{[n]}{j}$ the event N_S happens. Therefore,

$$Pr(\overline{Y_M}) = Pr\left(\bigvee_{j=1}^{p} \bigvee_{S \in \binom{[n]}{j}} N_S\right),$$

whence, we obtain:

$$Pr(Y_M) \geq 1 - \sum_{j=1}^{p} \binom{n}{j} \binom{j}{j - v_j + 1} \left(1 - (j - v_j + 1)\, x^{j-1}(1-x)\right)^m. \tag{3}$$

By the probabilistic method, there exists a $(p, \mathbf{v}, n)$-SUPER-SELECTOR of size $m^* = \text{argmin}_{m \geq 1} Pr(Y_M) > 0$. The rest of the proof will consist in showing that m^* satisfies the bound claimed.

Let us focus on the value c_j such that the j-th summand in (3) satisfies the following inequality

$$\binom{n}{j} \binom{j}{j - v_j + 1} \left(1 - (j - v_j + 1)\, x^{j-1}(1-x)\right)^{c_j j \log n/j} \leq 1/p \tag{4}$$

We shall use the following two inequalities

$$\left(1-(j-v_j+1)\,x^{j-1}(1-x)\right)^{c_j j\log(n/j)} \le \left(\frac{n}{j}\right)^{-\frac{(j-v_j+1)c_j j}{ep}} \tag{5}$$

$$\binom{n}{j}\binom{j}{j-v_j+1} \le n^j 2^{\frac{j}{2}} e^{\frac{3j}{2}} j^{-j} \tag{6}$$

By (5)-(6), we have that the left-hand-side of (4) can be upper bounded by

$$n^{j-\frac{c_j(j-v_j+1)j}{pe}} 2^{\frac{j}{2}} e^{\frac{3j}{2}} j^{-\left(j-\frac{c_j(j-v_j+1)j}{pe}\right)} = n^{j-\frac{c_j(j-v_j+1)j}{pe}} 2^{\frac{j}{2}} e^{\frac{3j}{2}} j^{-j+\frac{c_j(j-v_j+1)j}{pe}}, \tag{7}$$

Therefore, if we take $c_j = \frac{3pe}{(j-v_j+1)}$ we have that (7) can be further upper bounded with $n^{-2j}e^{2j}j^{2j}$ which is not larger than $1/p$ for all $n \ge 20$ and $n > p \ge j > 0$. Therefore, by taking

$$m = \max_{j=1,\ldots,p} c_j \log(n/j) = \max_{j=1,\ldots,p} \frac{3pej}{(j-v_j+1)} \log\frac{n}{j} \tag{8}$$

we can have each of the summands in (3) smaller than $1/p$, hence guaranteeing $Pr(Y_M) > 0$. By definition $m^* \le m$ which concludes the proof. □

The same analysis as above, tailored for a (p,k,n)-selector gives the following bounds.

Lemma 3. *For each $0 \le k < p < n$, there exists a (p,k,n)-selector of size*

$$m = \left(\log_2 \frac{e}{e-1+\frac{k}{p}}\right)^{-1} p\log\frac{n}{p}\,(1+o(1)) \le \frac{2p^2}{p-k+1}\log\frac{n}{p}\,(1+o(1)). \tag{9}$$

Moreover, there exists a (p,p,n)-selector of size $m = \dfrac{ep^2}{\log_2 e}\log(n/p)\,(1+o(1))$.

We can now combine the last two lemmas to obtain the main result of this section, providing an almost tight upper bound on the size of a SUPER-SELECTOR.

Theorem 3. *There exists a $(p,\mathbf{v},n)$-*SUPER-SELECTOR *of size*

$$m = O(\max_{j=1,\ldots,p} k_j \log(n/j)), \qquad \textit{where } k_j = \min\left\{\frac{3pej}{(j-v_j+1)}, \frac{ej^2}{\log_2 e}\right\}$$

Proof. Fix $k = \max\left\{j \mid \frac{3pej}{(j-v_j+1)} > \frac{ej^2}{\log_2 j}\right\}$. Let M_1 be a minimum size (k,k,n)-selector. In particular this is a $(k, <1,2,\ldots,k>,n)$-SUPER-SELECTOR hence *a fortiori* it is also a $(k,(v_1,\ldots,v_k),n)$-SUPER-SELECTOR.

Let M_2 be a minimum size $(p,(0,\ldots,0,v_{k+1},\ldots,v_p),n)$-SUPER-SELECTOR.

Let M be the binary matrix obtained by pasting together, one on top of the other, M_1 and M_2. It is not hard to see that M is a $(p,\mathbf{v},n)$-SUPER-SELECTOR. By Lemmas 3 and 2, M satisfies the desired bound. The proof is complete. □

Remark 3. Note that, if there exists a constant α such that $v_j \leq \alpha j$ for each $\sqrt{p} < j \leq p$, then the size of the SUPER-SELECTOR is $O(p \log \frac{n}{p})$, matching the information theoretic lower bound. Particular cases are given by instances where for each j, we have $v_j = f_j(j)$ for some function f_j such that $f_j(j) = o(j)$.

Deterministic construction. By using the method of the conditional expectations (see, e.g., [30]) we can derandomize the result of the previous section and provide a deterministic construction of the $(p, \mathbf{v}, n)$-SUPER-SELECTOR of Theorem 3 which is polynomial in n but exponential in the second parameter p. More precisely we obtain the following result, whose proof is deferred to full version of the paper.

Theorem 4. *There exists a deterministic* $O\left(p^3 n^{p+1} \log n\right)$ *construction of the* $(p, \mathbf{v}, n)$-SUPER-SELECTOR *given by Theorem 3.*

References

1. Alon, N., Asodi, V.: Tracing many users with almost no rate penalty. IEEE Trans. on Information Theory 53(1), 437–439 (2007)
2. Alon, N., Hod, R.: Optimal Monotone Encodings. IEEE Trans. on Information Theory 55(3), 1343–1353 (2009)
3. Balding, D.J., et al.: A comparative survey of non-adaptive pooling design. In: Speed, T.P., Waterman, M.S. (eds.) Genetic mapping and DNA sequencing, IMA Volumes in Mathematics and its Appl., pp. 133–154. Springer, Heidelberg (1996)
4. Chaudhuri, S., Radhakrishnan, J.: Deterministic restrictions in circuit complexity. In: Proc. of 28th STOC, pp. 30–36 (1996)
5. Cheng, Y., Du, D.Z.: New Constructions of One- and Two-Stage Pooling Designs. Journal of Computational Biology 15(2), 195–205 (2008)
6. Cheng, Y., Du, D.Z., Lin, G.: On the upper bounds of the minimum number of rows of disjunct matrices. Optimization Letters 3, 297–302 (2009)
7. Chlebus, B.S., Kowalski, D.R.: Almost Optimal Explicit Selectors. In: Liśkiewicz, M., Reischuk, R. (eds.) FCT 2005. LNCS, vol. 3623, pp. 270–280. Springer, Heidelberg (2005)
8. Cheraghchi, M.: Noise-resilient group testing: Limitations and constructions. In: Proc. of FCT 2009 (2009)
9. Chrobak, M., Gasieniec, L., Rytter, W.: Fast Broadcasting and Gossiping in Radio Networks. In: FOCS 2000, pp. 575–581 (2000)
10. Clementi, A.E.F., Monti, A., Silvestri, R.: Selective families, superimposed codes, and broadcasting on unknown radio networks. In: Proc. of Symp. on Discrete Algorithms (SODA 2001), pp. 709–718 (2001)
11. Cormode, G., Muthukrishnan, S.: Combinatorial Algorithms for Compressed Sensing. In: Flocchini, P., Gąsieniec, L. (eds.) SIROCCO 2006. LNCS, vol. 4056, pp. 280–294. Springer, Heidelberg (2006)
12. Cover, T.: Enumerative source encoding. IEEE Trans. Inf. Th. 19, 73–77 (1973)
13. Damaschke, P.: Adaptive versus Nonadaptive Attribute-Efficient Learning. In: STOC 1998, pp. 590–596 (1998)
14. De Bonis, A., Vaccaro, U.: Constructions of generalized superimposed codes with applications to group testing and conflict resolution in multiple access channels. Theoretical Computer Science 306, 223–243 (2003)

15. De Bonis, A., Gasieniec, L., Vaccaro, U.: Optimal Two-Stage Algorithms for Group Testing Problems. SIAM J. on Comp. 34(5), 1253–1270 (2005)
16. Du, D.Z., Hwang, F.K.: Pooling Design and Nonadaptive Group Testing. World Scientific, Singapore (2006)
17. D'yachkov, A.G., Rykov, V.V.: Bounds of the length of disjunct codes. Problems Control Inform. Theory 11, 7–13 (1982)
18. D'yachkov, A.G., Rykov, V.V., Rashad, A.M.: Superimposed distance codes. Problems Control Inform. Theory 18, 237–250 (1989)
19. Eppstein, D., Goodrich, M.T., Hirschberg, D.S.: Improved Combinatorial Group Testing Algorithms for Real-World Problem Sizes. SIAM J. on Comp. 36, 1360–1375 (2007)
20. Erdös, P., Frankl, P., Füredi, Z.: Families of finite sets in which no set is covered by the union of r others. Israel J. of Math. 51, 75–89 (1985)
21. Ganguly, S.: Data stream algorithms via expander graph. In: Hong, S.-H., Nagamochi, H., Fukunaga, T. (eds.) ISAAC 2008. LNCS, vol. 5369, pp. 52–63. Springer, Heidelberg (2008)
22. Gilbert, A.C., Iwen, M.A., Strauss, M.J.: Group Testing and Sparse Signal Recovery. In: 42nd Asilomar Conf. on Signals, Systems, and Computers, pp. 1059–1063 (2008)
23. Grebinsky, V., Kucherov, G.: Optimal Reconstruction of Graphs under the Additive Model. Algorithmica 28(1), 104–124 (2000)
24. Indyk, P.: Deterministic superimposed coding with application to pattern matching. In: Proc. of 39th FOCS 1997, pp. 127–136 (1997)
25. Indyk, P., Ngo, H.Q., Rudra, A.: Efficiently Decodable Non-adaptive Group Testing. In: Proc. of 20th SODA, pp. 1126–1142 (2010)
26. Kautz, W.H., Singleton, R.R.: Nonrandom binary superimposed codes. IEEE Trans. on Inform. Theory 10, 363–377 (1964)
27. Kumar, R., Rajagopalan, S., Sahai, A.: Coding constructions for blacklisting problems without computational assumptions. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 609–623. Springer, Heidelberg (1999)
28. Laczay, B., Ruszinkó, M.: Multiple User Tracing Codes. In: Proc. of ISIT 2006, pp. 1900–1904 (2006)
29. Linial, N.: Locality in distributed graph algorithms. SIAM J. on Computing 21, 311–312 (1992); Discrete Mathematics 162, 311-312 (1996)
30. Mitzenmacher, M., Upfal, E.: Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press, Cambridge (2005)
31. Moran, T., Naor, M., Segev, G.: Deterministic history-independent strategies for storing information on write-once memories. In: Arge, L., Cachin, C., Jurdziński, T., Tarlecki, A. (eds.) ICALP 2007. LNCS, vol. 4596, pp. 303–315. Springer, Heidelberg (2007)
32. Porat, B., Porat, E.: Exact and Approximate Pattern Matching in the Streaming Model. In: Proc. 50th FOCS, pp. 315–323 (2009)
33. Porat, E., Rothschild, A.: Explicit non-adaptive combinatorial group testing schemes. In: Aceto, L., Damgård, I., Goldberg, L.A., Halldórsson, M.M., Ingólfsdóttir, A., Walukiewicz, I. (eds.) ICALP 2008, Part I. LNCS, vol. 5125, pp. 748–759. Springer, Heidelberg (2008)
34. Clifford, R., Efremenko, K., Porat, E., Rothschild, A.: k-Mismatch with Don't Cares. In: Arge, L., Hoffmann, M., Welzl, E. (eds.) ESA 2007. LNCS, vol. 4698, pp. 151–162. Springer, Heidelberg (2007)
35. Ryabko, B.: Fast Enumerative Source Conding. In: Proc. of 1995 IEEE Intern. Symp. on Inf. Th., p. 395 (1995)
36. Wolf, J.: Born again group testing: Multiaccess Communications. IEEE Trans. Information Theory 31, 185–191 (1985)