# Improved List-Decodability and List-Recoverability of Reed–Solomon Codes via Tree Packings [Extended Abstract]

Zeyu Guo
*UT Austin*
*zguotcs@gmail.com*

Ray Li
*Stanford*
*rayyli@cs.stanford.edu*

Chong Shangguan
*Shandong University*
*theoreming@163.com*

Itzhak Tamo
*Tel Aviv University*
*zactamo@gmail.com*

Mary Wootters
*Stanford*
*marykw@stanford.edu*

2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS) | 978-1-6654-2055-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/FOCS52979.2021.00074

***Abstract*—This paper shows that there exist Reed–Solomon (RS) codes, over large finite fields, that are combinatorially list-decodable well beyond the Johnson radius, in fact almost achieving list-decoding capacity. In particular, we show that for any $\varepsilon \in (0,1]$ there exist RS codes with rate $\Omega(\frac{\varepsilon}{\log(1/\varepsilon)+1})$ that are list-decodable from radius of $1-\varepsilon$. We generalize this result to list-recovery, showing that there exist $(1-\varepsilon, \ell, O(\ell/\varepsilon))$-list-recoverable RS codes with rate $\Omega\left(\frac{\varepsilon}{\sqrt{\ell}(\log(1/\varepsilon)+1)}\right)$. Along the way we use our techniques to give a new proof of a result of Blackburn on optimal linear perfect hash matrices, and strengthen it to obtain a construction of strongly perfect hash matrices.**

**To derive the results in this paper we show a surprising connection of the above problems to graph theory, and in particular to the tree packing theorem of Nash-Williams and Tutte. We also state a new conjecture that generalizes the tree-packing theorem to hypergraphs, and show that if this conjecture holds, then there would exist RS codes that are *optimally* (non-asymptotically) list-decodable.**[1]

***Keywords*-Reed–Solomon codes; Nash-Williams–Tutte Theorem; Johnson radius; list decoding; list recovery; perfect hash matrix**

## I. Introduction

Reed–Solomon (RS) codes are a classical family of error correcting codes, ubiquitous in both theory and practice. To define an RS code, let $\mathbb{F}_q$ be the finite field of size $q$, and let $1 \le k < n \le q$. Fix $n$ distinct evaluation points $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{F}_q$. The $[n,k]$-*Reed–Solomon code* over $\mathbb{F}_q$ with evaluation points $(\alpha_1, \ldots, \alpha_n)$ is defined as the set

$$\left\{\left(f(\alpha_1), \ldots, f(\alpha_n)\right) \,:\, f \in \mathbb{F}_q[x],\ \deg(f) < k\right\}.$$

RS codes attain the optimal trade-off between *rate* and *distance*. The rate of a code $\mathcal{C} \subset \mathbb{F}_q^n$ is defined as $R = \log_q |\mathcal{C}|/n$. The rate is a number between 0 and 1, and the closer to 1 the better. The (relative) distance of a code $\mathcal{C} \subset \mathbb{F}_q^n$ is defined to be $\delta(\mathcal{C}) = \min_{c \neq c' \in \mathcal{C}} d(c,c')$, where $d(c,c') = |\{i \in [n] \,:\, c_i \neq c'_i\}|/n$ is relative Hamming distance. Again, the relative distance is a number between 0 and 1, and the closer to 1 the better. An $[n,k]$-RS code has rate $k/n$ and distance $(n-k+1)/n$, which is the best-possible trade-off, according to the Singleton bound [44].

Because RS codes attain this optimal trade-off (and also because they admit efficient algorithms), they have been well-studied since their introduction in the 1960's [39]. However, perhaps surprisingly, there is still much about them that we do not know. One notable example is their (combinatorial)[2] *list-decodability* and more generally their *list-recoverability*. We discuss list-decodability first, and discuss list-recoverability after that.

**List-decodability of RS codes:** List-decodability can be seen as a generalization of distance. For $\rho \in (0,1)$ and $L \ge 1$, we say that a code $\mathcal{C} \subset \mathbb{F}_q^n$ is $(\rho, L)$-list-decodable if for any $y \in \mathbb{F}_q^n$,

$$|\{c \in \mathcal{C} \,:\, d(c,y) \le \rho\}| \le L.$$

In particular, $(\rho, 1)$-list-decodability is the same as having distance greater than $2\rho$. List-decodability was introduced by Elias and Wozencraft in the 1950's [16], [47]. By now it is an important primitive in both coding theory and theoretical computer science more broadly. In general, larger *list sizes* (the parameter $L$) allow for a larger *list-decoding radius* (the parameter $\rho$). In this work, we will be interested in the case when $\rho = 1 - \varepsilon$ is large.

The list-decodability of Reed–Solomon codes is of interest for several reasons. First, both list-decodability and Reed–Solomon codes are central notions in coding theory, and the authors believe that question is interesting in its own right. Moreover, the list-decodability of Reed–Solomon codes has found applications in complexity theory and pseudorandomness [9], [45], [34].

Until recently, the best bounds available on the list-decodability of RS codes were bounds that hold generically for any code. The *Johnson bound* states that any code with minimum relative distance $\delta$ is $(1-\sqrt{1-\delta}, qn^2\delta)$-list-decodable over an alphabet of size $q$ ([29], see also [25, Theorem 7.3.3]). This implies that, for any $\varepsilon \in (0,1]$,

[1] A full version of this paper is available online at https://arxiv.org/abs/2011.04453.

[2] Throughout this paper, we will study *combinatorial* (rather than *algorithmic*) list-decodability.

2575-8454/21/$31.00 ©2021 IEEE
DOI 10.1109/FOCS52979.2021.00074

there are RS codes that are list-decodable up to radius $1-\varepsilon$ (with polynomial list sizes) that have rate $\Omega(\varepsilon^2)$. The celebrated Guruswami–Sudan algorithm [26] gives an efficient algorithm to list-decode RS codes up to the Johnson bound, but it breaks down at this point. Meanwhile, the *list-decoding capacity theorem* implies that no code (and in particular, no RS code) that is list-decodable up to radius $1-\varepsilon$ can have rate bounded above $\varepsilon$, unless the list sizes are exponential.

There have been several works over the past decade aimed at closing the gap between the Johnson bound (rate $\Theta(\varepsilon^2)$) and the list-decoding capacity theorem (rate $\Theta(\varepsilon)$). On the negative side, it is known that *some* RS codes (that is, some way of choosing the evaluation points $\alpha_1, \ldots, \alpha_n$), are not list-decodable substantially beyond the Johnson bound [4]. On the positive side, Rudra and Wootters [40] showed that a random choice of evaluation points will, with high probability, yield a code that is list-decodable up to radius $1-\varepsilon$ with rate $O\left(\frac{\varepsilon}{\log^5(1/\varepsilon)\log q}\right)$. Unfortunately, while the dependence on $\varepsilon$ in the rate is nearly optimal (the "correct" dependence should be linear in $\varepsilon$, according to the list-decoding capacity theorem), the $\log q$ term in the denominator means that the rate necessarily goes to zero as $n$ grows, as we must have $q \geq n$ for RS codes. Working in a different parameter regime, Shangguan and Tamo showed that over a large alphabet, there exist RS codes of rate larger than $1/9$ that can also be list-decoded beyond the Johnson bound (and in fact, optimally) [42]. However, this result only holds for small list sizes ($L = 2, 3$), and in particular, for such small list sizes one cannot hope to list-decode up to a radius $1-\varepsilon$ that approaches 1. Thus, there was still a substantial gap between capacity and the best known trade-offs for list-decoding RS codes.

**List-recoverability of RS codes:** The gap between capacity and the best known trade-offs for RS codes is even more pronounced for *list recovery,* a generalization of list decoding. We say that a code $\mathcal{C} \subset \mathbb{F}_q^n$ is $(\rho, \ell, L)$-list-recoverable if for any $S_1, S_2, \ldots, S_n \subset \mathbb{F}_q$ with $|S_i| = \ell$,

$$|\{c \in \mathcal{C} : d(c, S_1 \times S_2 \times \cdots \times S_n) \leq \rho\}| \leq L.$$

Here, we extend the definition of Hamming distance to sets by denoting

$$d(c, S_1 \times \cdots \times S_n) = \frac{1}{n}|\{i \in [n] : c_i \notin S_i\}|.$$

The parameter $\ell$ is called the *input list size.* List-decoding is the special case of list-recovery for $\ell = 1$. List-recovery first arose in the context of list-decoding (for example, the Guruswami–Sudan algorithm mentioned above is in fact a list-recovery algorithm), but has since found applications beyond that, for example in pseudorandomness [28] and algorithm design [14].

Both the Johnson bound and the list-decoding capacity theorem have analogs for list-recovery. The list-recovery Johnson bound [27] implies that there are RS codes of rate $\Omega(\varepsilon^2/\ell)$ that are list-recoverable up to radius $1-\varepsilon$ with input list size $\ell$ and polynomial output list size. However, the list-recovery capacity theorem implies that there are codes of rate $\Omega(\varepsilon)$ (with *no* dependence on $\ell$) that achieve the same guarantee, provided that the alphabet size $q$ is sufficiently large.

Thus the gap for list-recovery (between rate $\Theta(\varepsilon^2/\ell)$ and $\Theta(\varepsilon)$) is even larger than that for list-decoding, and in particular the dependence on $\ell$ becomes important. To the best of our knowledge, before our work there were *no* results known for RS codes that established list-recovery up to arbitrarily large radius $1-\varepsilon$ with a better dependence on $\ell$ than $1/\ell$.

**Motivating question:** Given this state of affairs, our motivating question is whether or not RS codes can be list-decoded or list-recovered up to radius $1-\varepsilon$ with rates $\Omega(\varepsilon)$ (in particular, with a linear dependence on $\varepsilon$ and no dependence on the alphabet size $q$ or the input list size $\ell$). As outlined below, we nearly resolve this question for list-decoding and make substantial progress for list-recovery.

**Subsequent work:** After this paper first appeared, and inspired by the techniques in this paper and in [42], Ferber, Kwan, and Sauermann showed that there exist $(1-\varepsilon, O(1/\varepsilon))$-list-decodable RS codes with rate $\Omega(\varepsilon)$ over a field size polynomial in the block length, improving our result for list-decoding [17]. In a very recent work, Goldberg, Shangguan, and Tamo further improved the rate of [17] by showing the existence of $(1-\varepsilon, O(1/\varepsilon))$-list-decodable RS codes with rate approaching $\frac{\varepsilon}{2-\varepsilon}$ [20]. See Section I-B for more details.

### A. Contributions

Our main result establishes the list-recoverability (and in particular, the list-decodability), of Reed–Solomon codes up to radius $1-\varepsilon$, representing a significant improvement over previous work. Our techniques build on the approach of [42]; the main new technical contribution is a novel connection between list-decoding RS codes and the Nash-Williams–Tutte theorem in graph theory, which may be of independent interest. We outline our contributions below.

**Existence of RS codes that are near-optimally list-decodable:** Our main theorem for list-decoding is as follows.

**Theorem I.1** (RS codes with near-optimal list-decoding)**.** *There is a constant $c \geq 1$ so that the following statement holds. For any $\varepsilon \in (0, 1]$ and any sufficiently large $n$, there exist RS codes of rate $R \geq \frac{\varepsilon}{c(\log(1/\varepsilon)+1)}$ over a large enough*

*finite field (as a function of $n$ and $\varepsilon$), that are $(1-\varepsilon, c/\varepsilon)$-list-decodable.*

As discussed above, Theorem I.1 is stronger than the result of Rudra and Wootters [40], in that the result of [40] requires that the rate tend to zero as $n$ grows, while ours holds for constant-rate codes. On the other hand, our result requires the field size $q$ to be quite large (see Table I), which [40] did not require.

Our result also differs from the result of Shangguan and Tamo [42] discussed above. Because that work focuses on small list sizes, it does not apply to list-decoding radii approaching 1. In contrast, we are able to list-decode up to radius $1-\varepsilon$. We note that [42] is able to show that RS codes are exactly optimal, while we are off by logarithmic factors. Both our work and that of [42] require large field sizes.

**Generalization to list-recovery:** Theorem I.1 follows from a more general result about list-recovery. Our main result is the following (see Theorem 5.1 in [21] for a more detailed version).

**Theorem I.2** (RS codes with list-recovery beyond the Johnson bound)**.** *There is a constant $c \geq 1$ such that the following statement holds. For any $\varepsilon \in (0,1]$, any positive integer $\ell$, and any sufficiently large $n$, there exist RS codes with rate $R \geq \frac{\varepsilon}{c\sqrt{\ell}(\log(1/\varepsilon)+1)}$ over a large enough (as a function of $n$, $\varepsilon$, and $\ell$) finite field, that are $(1-\varepsilon, \ell, c\ell/\varepsilon)$-list-recoverable.*

Theorem I.2 establishes list-recoverability for RS codes well beyond the Johnson bound, and in particular breaks the $1/\ell$ barrier. To the best of our knowledge, this is the first result to do so for radius arbitrarily close to 1, although we note that work of Lund and Potukuchi achieved a similar rate for small error radius [34]. We discuss related work below in Section I-B and summarize quantitative results in Table I.

**Applications to perfect hashing:** Our techniques also have an application to the construction of *strongly perfect hash matrices*, as detailed below. Given a matrix and a set $S$ of its columns, a row is said to *separate* $S$ if, restricted to this row, these columns have distinct values. For a positive integer $t$, a matrix is said to be a *$t$-perfect hash matrix* if any set of $t$ distinct columns of the matrix is separated by at least one row. Perfect hash matrices were introduced by Mehlhorn [35] in 1984 for database management, and since then they have found various applications in cryptography [7], circuit design [37], and the design of deterministic analogs of probabilistic algorithms [3].

Let $\mathrm{PHF}(n, m, q, t)$ denote a $q$-ary $t$-perfect hash matrix with $n$ rows and $m$ columns. Given $m, q, t$, determining the minimal $n$ such that there exists a $\mathrm{PHF}(n, m, q, t)$ is one of the major open questions in this field, and has received considerable attention (see, e.g., [8], [6], [41]). For any integers $t \geq 2,\ k \geq 2$, and sufficiently large prime power $q$, using tools from linear algebra Blackburn [8] constructed a $\mathrm{PHF}(k(t-1), q^k, q, t)$, which remains the best-known construction for such parameters so far.

Constructing perfect hash matrices is related to list-recovery and list-decoding. Indeed, if the columns of our matrix are codewords, then the matrix is a $t$-perfect hash matrix if and only if the code is $(0, t-1, t-1)$-list-recoverable. On the way to proving our main result on list-recovery, we prove a theorem (see Theorem I.9 below) that gives very precise bounds, but only in a restricted setting. While this setting is too restrictive to immediately yield results on list-recovery in general, it turns out to be enough to say something interesting about perfect $t$-hash matrices. In particular, we are able to recover Blackburn's result, and extend it to a generalization of perfect hashing where every set of $t$ columns needs to be separated not just by one row but by many rows.

**Theorem I.3.** *Given integers $1 \leq k < n$ and $t \geq 3$, for a sufficiently large prime power $q$, there exists an $n \times q^k$ matrix, defined on the alphabet $\mathbb{F}_q$, such that any set of $t$ columns is separated by at least $n - k(t-1) + 1$ rows.*

We call a matrix with the property given by Theorem I.3 a *strongly $t$-perfect hash matrix*; this can be viewed as an "error-resilient" version of perfect hash matrices. Strongly perfect hash matrices were first introduced by the third and fourth authors of this paper for $t = 3$, with a slightly different definition [43]. Indeed, Lemma 25 of [43] implies the $t = 3$ case of Theorem I.3, but it breaks down at that point. We overcome this barrier, and construct strongly $t$-perfect hash matrices for all integers $t \geq 3$. The main ingredient in our proof is a surprising connection from strongly perfect hashing to graph theory (see Lemma I.5 and the discussion after it for the details).

Generalizing a definition of [8] (with a slightly different terminology), we say that an $n \times q^k$ matrix $M$ is called *linear* if it is defined over the field $\mathbb{F}_q$ and has the form $M = PQ$, where $P$ is an $n \times k$ coefficient matrix and $Q$ is the $k \times q^k$ matrix whose columns are formed by the $q^k$ distinct vectors of $\mathbb{F}_q^k$.

With this terminology, we will prove the following proposition, which generalizes a result of [8] (see Theorem 4 of [8]).

**Proposition I.4.** *If a linear $n \times q^k$ matrix separates any set of $t$ columns by at least $r$ rows, then $r \leq n - k(t-1) + 1$.*

Proposition I.4 implies that the bound in Theorem I.3 is tight, at least for linear constructions.

**A new connection to the Nash-Williams–Tutte theorem, and a new hypergraph Nash-Williams–Tutte conjecture:**

Table I
PRIOR WORK ON LIST-DECODING AND LIST-RECOVERY OF RS CODES. ABOVE, $C$ REFERS TO AN ABSOLUTE CONSTANT. THE "CAPACITY" RESULTS REFER TO THE LIST-DECODING AND LIST-RECOVERY CAPACITY THEOREMS, RESPECTIVELY, AND ARE IMPOSSIBILITY RESULTS. ABOVE, WE ASSUME THAT $q \geq n$ AND THAT $n \to \infty$ IS GROWING RELATIVE TO $1/\varepsilon$ AND $\ell$, AND THAT $n$ IS SUFFICIENTLY LARGE.

| | Radius $\rho$ | List size $L$ | Rate $R$ | Field size $q$ |
|---|---|---|---|---|
| **List-Decoding:** | | | | |
| Capacity | $1-\varepsilon$ | - | $\leq \varepsilon$ | - |
| Johnson bound | $1-\varepsilon$ | $\mathrm{poly}(n)$ | $C\varepsilon^2$ | $q \geq n$ |
| [40] | $1-\varepsilon$ | $C/\varepsilon$ | $\frac{C\varepsilon}{\log^5(1/\varepsilon)\log(q)}$ | $q \geq Cn\log^C(n/\varepsilon)/\varepsilon$ |
| [42] | $\frac{L}{L+1}(1-R)$ | $L = 2, 3$ | $R$ | $q = 2^{Cn}$ |
| This work (Thm. I.1) | $1-\varepsilon$ | $C/\varepsilon$ | $\frac{C\varepsilon}{\log(1/\varepsilon)}$ | $q = \left(\frac{1}{\varepsilon}\right)^{Cn}$ |
| **List-Recovery:** | | | | |
| Capacity | $1-\varepsilon$ | - | $\leq \varepsilon$ | - |
| Johnson bound | $1-\varepsilon$ | $\mathrm{poly}(n)$ | $\frac{C\varepsilon^2}{\ell}$ | $q \geq n$ |
| [34] | $\rho \leq 1-1/\sqrt{2}$ | $C\ell$ | $\frac{C}{\sqrt{\ell}\cdot\log q}$ | $q \geq Cn\sqrt{\ell}\cdot\log n$ |
| This work (Thm. I.2) | $1-\varepsilon$ | $\frac{C\ell}{\varepsilon}$ | $\frac{C\varepsilon}{\sqrt{\ell}\cdot\log(1/\varepsilon)}$ | $q = \left(\frac{\ell}{\varepsilon}\right)^{Cn}$ |

In order to derive our results, we build on the framework of [42]. That work developed a framework to view the list-decodability of Reed–Solomon codes in terms of the singularity of *intersection matrices* (which we define in Section II). The main new technical contribution of our work is to connect the singularity of these matrices to tree-packings in particular graphs. This connection allows us to use the Nash-Williams–Tutte theorem from graph theory to obtain our results. The Nash-Williams–Tutte theorem gives sufficient conditions for the existence of a large *tree packing* (that is, a collection of pairwise edge-disjoint spanning trees) in a graph.

**Lemma I.5** (Nash-Williams [36], Tutte [46], see also Theorem 2.4.1 of [13])**.** *A multigraph contains $k$ edge-disjoint spanning trees if and only if for every partition $\mathcal{P}$ of its vertex set it has at least $(|\mathcal{P}|-1)k$ cross-edges. Here an edge is called a cross-edge for $\mathcal{P}$ if its two endpoints are in different members of $\mathcal{P}$.*

Lemma I.5 is of particular importance for the proofs of the main results of this paper, e.g., Theorems I.1, I.2, and I.3. We think that this connection with graph theory is a contribution in its own right, and it is our hope that it will lead to further improvements to our results on Reed–Solomon codes. In particular, we hope that it will help establish the following conjecture of [42]:

**Conjecture I.6** (Conjecture 1.5 of [42])**.** *For any $\varepsilon > 0$ and integers $1 \leq k < n$ with $\varepsilon n \in \mathbb{Z}$, there exist RS codes with rate $R = \frac{k}{n}$ over a large enough (as a function of $n$ and $\varepsilon$) finite field, that are list-decodable from radius $1-R-\varepsilon$ and list size at most $\lceil \frac{1-R-\varepsilon}{\varepsilon} \rceil$.*

Conjecture I.6 is stronger than our Theorem I.1 about list-decoding. In particular, our theorem is near-optimal, but it is interesting mostly in the low-rate/high-noise parameter regime. In contrast, Conjecture I.6 conjectures that there exist *exactly* optimal RS codes, in any parameter regime.

To encourage others to use our new connection and make progress on Conjecture I.6, we propose a method of attack in Section 6 of [21]. This outline exploits our connection to the Nash-Williams–Tutte theorem, and proceeds via a conjectured generalization of the Nash-Williams–Tutte theorem to hypergraphs (see Conjecture III.1 below): we show that establishing this hypergraph conjecture would in fact establish Conjecture I.6.

**Theorem I.7.** *Conjecture III.1 implies Conjecture I.8 and thus Conjecture I.6.*

As further evidence of the viability of this approach, this quantitative relaxation implies a second proof of our main list-decoding result, Theorem I.1, and we also sketch this proof in Section 6 of [21].[3]

### B. Related Work

We briefly review related work. See Table I for a quantitative comparison to prior work.

**List-decoding of RS codes:** Ever since the Guruswami–Sudan algorithm [26], which efficiently list-decodes RS codes up to the Johnson bound, it has been open to understand the extent to which RS codes are list-decodable *beyond* the Johnson bound, and in particular if there are RS codes that are list-decodable all the way up to the list-decoding capacity theorem, matching the performance of completely random codes. There have been negative results that show that *some* RS codes are not list-decodable to

[3]This second proof does not immediately establish list-recoverability, which is why we focus on our first proof.

capacity [4], and others that show that even if they were, in some parameter regimes we are unlikely to find an efficient list-decoding algorithm [11]. The work of Rudra and Wootters, mentioned above, showed that for any code with suitably good distance, a random puncturing of that code was likely to be near-optimally list-decodable; this implies that an RS code with random evaluation points is likely to be list-decodable. Unfortunately, as discussed above, this result requires a constant alphabet size $q$ in order to yield a constant-rate code, while RS codes necessarily have $q \geq n$.

Recently, Shangguan and Tamo [42] studied the list-decodability of RS codes in a different parameter regime, namely when the list size $L$ is very small, either 2 or 3. They were able to get extremely precise bounds on the rate (showing that there are RS codes that are exactly optimal), but unfortunately for such small list sizes, it is impossible for any code to be list-decodable up to radius $1 - \varepsilon$ for small $\varepsilon$, which is our parameter regime of interest. Unlike the approach of [40], which applies to random puncturings of any code, the work of [42] targeted RS codes specifically and developed an approach via studying *intersection matrices*. The reason that their approach stopped at $L = 3$ was the difficulty of analyzing these intersection matrices. We build on their approach and use techniques from graph theory—in particular, the Nash-Williams–Tutte theorem—to analyze the relevant intersection matrices beyond what [42] were able to do. We discuss our approach more below in Section I-C.

**Subsequent work on list-decoding of RS codes:** After our work first appeared, and inspired by our approach, Ferber, Kwan, and Sauermann [17] gave a beautiful proof establishing the existence of RS codes with rate $\Omega(\varepsilon)$ that are list-decodable from radius $1 - \varepsilon$ with list size $O(1/\varepsilon)$, over a polynomially (in the code's length) large finite field.[4] Compared with our result on the list-decodability of RS codes, their result removes the logarithmic factor in $1/\varepsilon$, and allows for smaller alphabet sizes; additionally, their proof is much shorter. In further follow-up work, Goldberg, Shangguan, and Tamo [20] further improved the rate from $\Omega(\varepsilon)$ to a rate approaching $\frac{\varepsilon}{2-\varepsilon}$.

However, we believe that there are still some advantages to our approach (beyond inspiring that of [17] and [20]). First, the result of [17] does not apply to list-recovery, and while [20] does apply to list-recovery, they do not surpass the $1/\ell$ barrier in the rate. Second, neither [17] nor [20] fully resolve Conjecture I.6 about optimal list-decodability of RS codes. We believe that the framework and tools developed in this paper together with the hypergraph Nash-Williams–Tutte conjecture (see [21]) provide a plausible attack method to resolve Conjecture I.6.

[4]In fact, they show something more general: if one begins with any code of sufficiently large distance over a sufficiently large alphabet, and randomly punctures it to rate $\Omega(\varepsilon)$, the resulting code is with high probability $(1 - \varepsilon, O(1/\varepsilon))$ list-decodable.

**List-recovery of RS codes:** While the Guruswami–Sudan algorithm is in fact a list-recovery algorithm, much less was known about the list-recovery of RS codes beyond the Johnson bound than was known about list-decoding. (There is a natural extension of the Johnson bound for list-recovery, see [27]; for RS codes, it implies that an RS code of rate about $\varepsilon^2/\ell$ is list-recoverable up to radius $1 - \varepsilon$ with input list sizes $\ell$ and polynomial output list size). As with list-decoding, it is known that *some* RS codes are not list-recoverable beyond the Johnson bound [22]. However, much less was known on the positive front. In particular, neither of the works [40], [42] discussed above work for list-recovery. In a recent work, Lund and Potuchuki [34] have proved an analogous statement to that of [40]: any code of decent distance, when randomly punctured to an appropriate length, yields with high probability a good list-recoverable code. This implies the existence of RS codes that are list-recoverable beyond the Johnson bound. However, in [34] there is again a dependence on $\log(q)$ in the rate bound, meaning that for RS codes, the rate must be sub-constant. Further, the work of [34] only applies up to radius $\rho = 1 - 1/\sqrt{2}$, and in particular does not apply to radii $\rho = 1 - \varepsilon$, as we study in this work. Our results also work in the constant-$\rho$ setting of [34], and in that regime we show that RS codes of rate $\Omega(1/\sqrt{\ell})$ are $(\rho, \ell, O(\ell))$ list-recoverable, which improves over the result of [34] by a factor of $\log q$ in the rate. However, we do require the field size to be much larger than that is required by [34] (see Table I).

**Subsequent work on list-recovery of RS codes:** The recent work of Goldberg, Shangguan, and Tamo [20] mentioned above builds on [17], and shows that there are RS codes of rate approaching $\frac{\varepsilon}{1+\ell-\varepsilon}$ that are $(1 - \varepsilon, \ell, L_{\varepsilon,\ell})$-list-recoverable, for a constant $L_{\varepsilon,\ell}$ that depends only on $\varepsilon$ and $\ell$. Compared to our work, while [20] improves the dependence on $\varepsilon$ in the rate by a factor of $\log(1/\varepsilon)$, it has a worse dependence on $\ell$, and in particular does not break the $1/\ell$ barrier that is present in the Johnson bound.

**List-decoding and list-recovery of RS-like codes:** There are constructions—for example, of *folded RS codes* and *univariate multiplicity codes* [24], [23], [31], [30]—of codes that are based on RS codes and that are known to achieve list-decoding (and list-recovery) capacity, with efficient algorithms. Our goal in this work is to study Reed–Solomon codes themselves.

**Perfect hash matrices and strongly perfect hash matrices:** Perfect hash matrices have been studied extensively since the 1980s. There are two parameter regimes that are studied. The first is when the alphabet size $q$ is constant and

the number of rows tends to infinity [38], [19], [33], [32], [48]. The second is when the number of rows is viewed as a constant, while $q$ may tend to infinity [8], [6], [41]. In both cases the strength $t$ of a perfect hash matrix is a constant. Our work studies the second case; as mentioned above, Blackburn [6] gave an optimal construction for linear hash matrices in this parameter regime, and as a special case we obtain a second proof of Blackburn's result.

The study of strongly perfect hash matrices is relatively new [43]. The thesis [15] collected some recent results on a closely related topic. However, the parameters considered there are quite different from those in our paper, and to the best of our knowledge, our construction is the best known in the parameter regime we consider. Another related notion called *balanced hashing* was introduced in [1], [2], where, with our notation, any set of $t$ columns of a matrix needs to be separated by at least $a_1$ and at most $a_2$ rows, for some integers $a_1 \le a_2$. Note that in our setting, we want every set of $t$ columns to be separated by as many rows as possible, while in the setting of balanced hashing it cannot exceed the threshold $a_2$; thus, the two settings are incomparable.

### C. Technical Overview

**Intersection matrices:** Our approach is centered around *intersection matrices,* introduced in [42]. Intersection matrices and their nonsingularity are defined formally below in Definition II.2, but we give a brief informal introduction here. A $t$-wise intersection matrix, $M$, is defined by a collection of sets $I_1, I_2, \ldots, I_t \subseteq [n]$, and has entries that are monomials in $\mathbb{F}_q[x_1, x_2, \ldots, x_n]$. It was shown in [42] that if there is a counter-example to the list-decodability of a Reed–Solomon code with evaluation points $(\alpha_1, \ldots, \alpha_n)$—that is, if there exist polynomials $f_1, f_2, \ldots, f_{L+1}$ that all agree with some other polynomial $g : \mathbb{F}_q \to \mathbb{F}_q$ at many points $\alpha_i$—then there is a $(L+1)$-wise intersection matrix that becomes singular when $\alpha_i$ is plugged in for $x_i$ for all $i \in [n]$.

The set-up (both the definition of an intersection matrix and the connection to list-decoding) is most easily explained by an example. Suppose that we are interested in list-decoding for $L = 3$, and suppose that we are interested in a RS code with evaluation points $\alpha_1, \alpha_2, \ldots, \alpha_n$. Let $f_1, f_2, f_3, f_4$ and $g$ be a counter-example to list-decoding, as above, and for $1 \le j \le 4$, let $I_j = \{i \in [n] : f_j(\alpha_i) = g(\alpha_i)\}$. Now consider the product shown in Figure 1 (see the caption for notation).

An inspection of Figure 1 shows that the matrix-vector product depicted is zero. Indeed, the top part is zero for any choice of the $f_i$, and the bottom part is zero since $f_i$ and $f_j$ are assumed to agree on $\{\alpha_s : s \in I_i \cap I_j\}$. The matrix shown is the 4-wise intersection matrix for the sets $I_1, I_2, I_3, I_4$, evaluated at $\alpha_1, \ldots, \alpha_n$. If the $f_i$'s agree too much with the function $g$ (i.e., if they are a counter-example to list-decodability for some given radius), then the sets $I_i \cap I_j$ are going to be larger, and this matrix will have more rows. In particular, the more the $f_i$'s agree with $g$, the harder it is for this matrix to be singular. Intuitively, this sets us up for a proof by contradiction: if $f_1, f_2, f_3, f_4$ agree too much with $g$, then this matrix is nonsingular (at least for a non-pathological choice of $\alpha_i$'s); but Figure 1 displays a kernel vector!

A $t$-wise intersection matrix (for sets $I_1, \ldots, I_t$) generalizes a 4-wise intersection matrix shown in Figure 1. The bottom part looks exactly the same—a block-diagonal matrix with Vandermonde blocks—and the top part is an appropriate generalization that causes the analogous $k \cdot \binom{t}{2}$-long vector corresponding to the $f_i$'s to vanish.

**A conjecture about $t$-wise intersection matrices:** With the motivation in Figure 1, the strategy of [42] was to study $t$-wise intersection matrices $M$ for $t = L + 1$, and to show that for every appropriate choice of $I_1, \ldots, I_t$, the polynomial $\det(M) \in \mathbb{F}_q[x_1, x_2, \ldots, x_n]$ is not identically zero. The list-decodability of RS codes would then follow from the DeMillo–Lipton–Schwartz—Zippel lemma along with a counting argument. In particular, they made the following conjecture, and showed that it implies Conjecture I.6 about list-decoding. Below, the *weight* of a family of subsets $I_1, \ldots, I_t$ of $[n]$ is defined to be

$$\mathrm{wt}(I_1, \ldots, I_t) = \sum_{i=1}^{t} |I_i| - \left| \bigcup_{i=1}^{t} I_i \right|,$$

and for a set $J$ of indices, we use the shorthand $\mathrm{wt}(I_J) := \mathrm{wt}(I_j : j \in J)$.

**Conjecture I.8** (Conjecture 5.7 of [42])**.** *Let $t \ge 3$ be an integer and $I_1, \ldots, I_t \subseteq [n]$ be subsets satisfying*

(i) $\mathrm{wt}(I_J) \le (|J| - 1)k$ *for all nonempty* $J \subseteq [t]$,
(ii) *Equality holds for* $J = [t]$, *i.e.,* $\mathrm{wt}(I_{[t]}) = (t-1)k$.

*Then the $t$-wise intersection matrix $M_{k,(I_1,\ldots,I_t)}$ is nonsingular over any finite field.*

The conditions (i) and (ii) above turn out to be the right way of quantifying "the $f_i$'s agree enough with $g$." That is, if the $f_i$'s agree too much with $g$ (in the sense of going beyond Conjecture I.6 about list-decoding), then it is possible to find sets $I_j$ so that (i) and (ii) hold.

Unfortunately, the work of [42] was only able to establish Conjecture I.8 for $t = 3, 4$ (corresponding to $L = 2, 3$), and it seemed challenging to extend their techniques directly to much larger values of $L$.

**Establishing the conjecture under an additional assumption, and using that to establish our main results:** In this work, we use a novel connection to the Nash-Williams–Tutte theorem, which establishes the existence of pairwise edge-disjoint spanning trees in a graph, to extend the results of [42] to larger $L$, at the cost of an additional assumption.

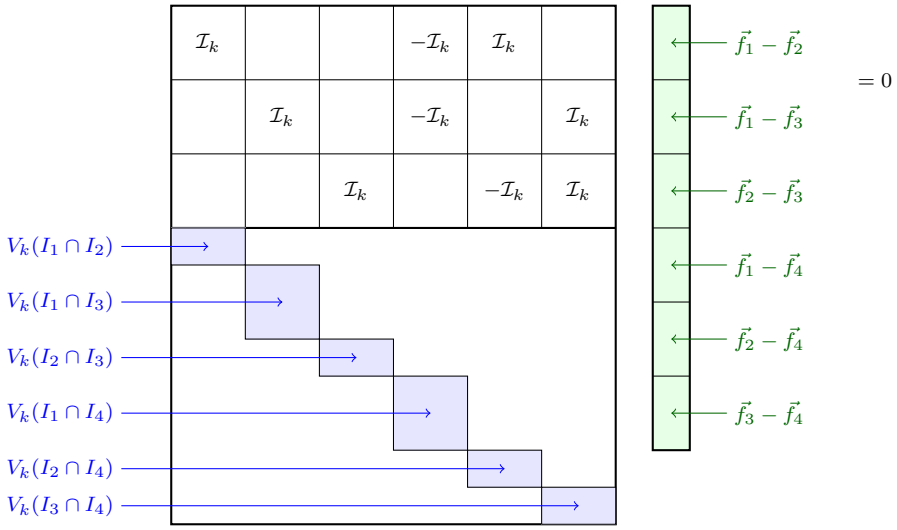


Figure 1. Let $f_1, f_2, f_3, f_4 \in \mathbb{F}_q[x]$ have degree $k-1$ and suppose that $I_j = \{s : f_j(\alpha_s) = g(\alpha_s)\}$. (In particular, $f_i$ and $f_j$ agree on $I_i \cap I_j$). Then the matrix-vector product depicted above is zero, where the vector $\vec{f_i}$ refers to the $k$ coefficients of the polynomial $f_i$, and the $j$-th coordinate of this vector is the coefficient of $x^{j-1}$ in $f$. Here, $V_k(I_i \cap I_j) \in \mathbb{F}_q^{|I_i \cap I_j| \times k}$ denotes the Vandermonde matrix whose rows are $[\alpha_s^0, \alpha_s^1, \ldots, \alpha_s^{k-1}]$ for $s \in I_i \cap I_j$. The notation $\mathcal{I}_k$ denotes the $k \times k$ identity matrix.

More precisely, although not able to prove Conjecture I.8 in its full generality, we are able to prove a special case of it, as stated below.

**Theorem I.9.** *Let $t \geq 2$ be an integer and $I_1, \ldots, I_t \subseteq [n]$ be subsets satisfying*

(i) $I_i \cap I_j \cap I_l = \emptyset$ *for all* $1 \leq i < j < l \leq t$;
(ii) $\mathrm{wt}(I_J) \leq (|J| - 1)k$ *for all nonempty* $J \subseteq [t]$;
(iii) $\mathrm{wt}(I_{[t]}) = (t-1)k$.

*Then, the $t$-wise intersection matrix $M_{k,(I_1,\ldots,I_t)}$ is nonsingular over any field.*

Clearly, this theorem stops short of Conjecture I.8, due to the assumption that $I_i \cap I_j \cap I_\ell = \emptyset$. However, we will build on this statement to prove our main theorem about list-recovery (Theorem I.2), and moreover this is already enough to prove our result on the existence of strongly perfect hash matrices (Theorem I.3). Theorem I.9 follows from an interesting application of Lemma I.5. We only sketch the main ideas here, leaving the details to [21].

Briefly, we consider each term in the expression

$$\det(M) = \sum_{\sigma \in S_n} (-1)^{\mathrm{sgn}(\sigma)} \prod_{i=1}^{n} M_{i,\sigma(i)}.$$

We show that $\prod_{i=1}^{n} M_{i,\sigma(i)}$ is a nonzero monomial in $x_1, \ldots, x_n$ if and only if $\sigma$ picks out a tree packing of a graph[5] that is determined by the sets $I_1, \ldots, I_t$. It turns out that the requirements of (i) and (ii) in Conjecture I.8 translate exactly into the requirements needed to apply the Nash-Williams–Tutte theorem to this graph. Thus, if (i) and (ii) hold, then there exists a tree packing in this graph and hence a nonzero term in $\det(M)$.

If the sets $I_i \cap I_j$ and $I_{i'} \cap I_{j'}$ that appear in the lower part of the $t$-wise intersection matrix do not intersect (that is, if there are no three-wise intersections among the sets $I_j$), then the reasoning above is enough to establish the conclusion of Conjecture I.8, because all of the terms that appear in the expansion of the determinant are distinct monomials, and they cannot cancel. This is why Theorem I.9 has this assumption.

While Theorem I.9 is not strong enough to immediately establish results for list-decoding or list-recovery (indeed, there is no reason that there should not be three-wise intersections for the polynomials $f_i$ discussed above), it *is* enough for our application to perfect hash matrices (see Section 4 of [21]).

In order to apply Theorem I.9 to list-decoding, we back off from Conjecture I.8 a bit. First, we allow a factor of $\Theta(\log t)$ slack on the right-hand sides of (i) and (ii). Second, rather than showing that the $t$-wise intersection matrix $M_{k,(I_1,\ldots,I_t)}$ is nonsingular, we show that there exists a $t'$-wise intersection matrix that is nonsingular for some $t' < t$. Following the connection of [42] illustrated in Figure 1, this turns out to be enough to establish our main theorem on list-decoding/recovery.

We choose this smaller intersection matrix by carefully choosing a random subset $J$ of $[t]$. By greedily removing elements from the sets $\{I_j : j \in J\}$, we can obtain subsets $I'_j \subset I_j$ with empty three-wise intersections $I'_j \cap I'_{j'} \cap I'_{j''} = \emptyset$. Furthermore, by the careful random choice of $J$, and since

[5]Throughout this paper, a tree packing of a graph $G$ means a collection of pairwise edge-disjoint spanning trees of $G$.

we allowed a $\Theta(\log t)$ slack in the initial weight bounds, we can show this step does not delete too many elements. This is the key step of our proof. Using some of the sets $\{I_j : j \in J\}$, we can find a smaller intersection matrix obeying the setup of Conjecture I.8 with the additional guarantee that all three-wise intersections are empty. We provide a more detailed summary of the proof in Section 5.1 of [21].

**Another avenue to list-decoding, a hypergraph Nash-Williams–Tutte conjecture:** Extending our connection of list-decoding RS codes to the Nash-Williams–Tutte theorem, we show that a suitable *hypergraph* generalization of the Nash-Williams–Tutte theorem would imply Conjecture I.8 about the nonsingularity of intersection matrices, without any need for an additional assumption about three-wise intersections of the sets $I_j$.

We conjecture that such a generalization is true, and we state it in Section III as Conjecture III.1. We show that if Conjecture III.1 were true, it would imply Conjecture I.8, on the nonsingularity of intersection matrices. This in turn would imply Conjecture I.6, establishing the existence of RS codes with optimal list-decodability. This suggests a plan of attack towards Conjecture I.6.

While we are unable to establish this challenging conjecture in full, we give some evidence for it. First, we show that the "easy part" of the conjecture follows from the Nash-Williams–Tutte theorem. Second, we observe that a quantitative relaxation of the conjecture follows from known results on Steiner tree packings [12] and disjoint bases of polymatroids [10]. This relaxation can be combined with the connection of hypergraph packings and intersection matrices, and the connection between intersection matrices and list decoding RS codes, to give a second proof of Theorem I.1, that there are *near*-optimally list-decodable RS codes.

In addition to implying the optimal list-decodability of RS codes, Conjecture III.1 may be of independent interest. A hypergraph generalization of Nash-Williams–Tutte is known for *partition-connected* hypergraphs [18] (see Section III for definition), a well studied notion. However, for a different notion called *weak-partition-connectivity*, less seems to be known, and Conjecture III.1 poses a Nash-Williams–Tutte generalization for weakly-partition-connected hypergraphs.

**Organization.** A graphical overview of our results can be found in Figure 2. In Section II we will give the formal definition of intersection matrices. In Section III we will introduce our conjectured hypergraph version of the Nash-Williams–Tutte theorem. The proofs of all the results mentioned above can be found in [21].

### D. Future Directions and Open Questions

In this work, we have shown the existence of near-optimally list-decodable RS codes in the large-radius parameter regime. To do this, we have established a connection between the intersection matrix approach of [42] and tree packings. Along the way, we also developed applications to the construction of strongly perfect hash matrices, and we have introduced a new hypergraph version of the Nash-Williams–Tutte theorem. We highlight a few questions that remain open.

**Can RS codes *exactly* achieve list-decoding capacity?** In spite of the results and tools developed in this paper, we were not able to prove Conjecture I.6. We hope that the avenue of attack discussed in Section III will be able to finish the job. We note that the analogous question regarding the limits of list-recoverability of RS codes also remains open.

**Efficient list-decoding of RS codes?** We remark that, using a simple idea from [42] one can convert each of the existence results of RS codes reported in this paper into an explicit code construction, although over a much larger field size. Hence, given such an explicit code construction, is it possible to decode it efficiently up to its guaranteed list-decoding radius? A similar question can be asked for list-recoverability. We note that [11], which shows that decoding RS codes much beyond the Johnson bound is likely hard in certain parameter regimes, does not apply to our parameter regime when the field size is large.

**Generalizing the Nash-Williams–Tutte theorem to hypergraphs:** In an attempt to resolve Conjecture I.6, we present Conjecture III.1, a new graph-theoretic conjecture, which can be viewed as a generalization of the Nash-Williams–Tutte theorem to hypergraphs. In addition to being interesting on its own, resolving this conjecture would imply the existence of optimally list-decodable RS codes.

## II. Intersection Matrices

The main goal of this section is to present the definition of $t$-wise intersection matrices over an arbitrary field $\mathbb{F}$.

Let $\mathbb{N}^+ = \{1, 2, \dots\}$ and $[n] = \{1, 2, \dots, n\}$ for $n \in \mathbb{N}^+$. Denote by $\log x$ the base-2 logarithm of $x$. For a finite set $X$ and an integer $1 \le k \le |X|$, let $\binom{X}{k} = \{A \subseteq X : |A| = k\}$ be the family of all $k$-subsets of $X$. For an integer $t \ge 3$, we define the following lexicographic order on $\binom{[t]}{2}$. For distinct $S_1, S_2 \in \binom{[t]}{2}$, $S_1 < S_2$ if and only if $\max(S_1) < \max(S_2)$ or $\max(S_1) = \max(S_2)$ and $\min(S_1) < \min(S_2)$. For a partition $\mathcal{P}$ of $X$, let $|\mathcal{P}|$ denote the number of parts of $\mathcal{P}$. In the remaining part of this paper, assume that $n, k$ are integers satisfying $1 \le k < n$.

We view a polynomial $f \in \mathbb{F}_q[x]$ of degree at most $k-1$ as a vector of length $k$ defined by its $k$ coefficients, where for $1 \le i \le k$, the $i$-th coordinate of this vector is the coefficient of $x^{i-1}$ in $f$. By abuse of notation that vector is also denoted by $f$.

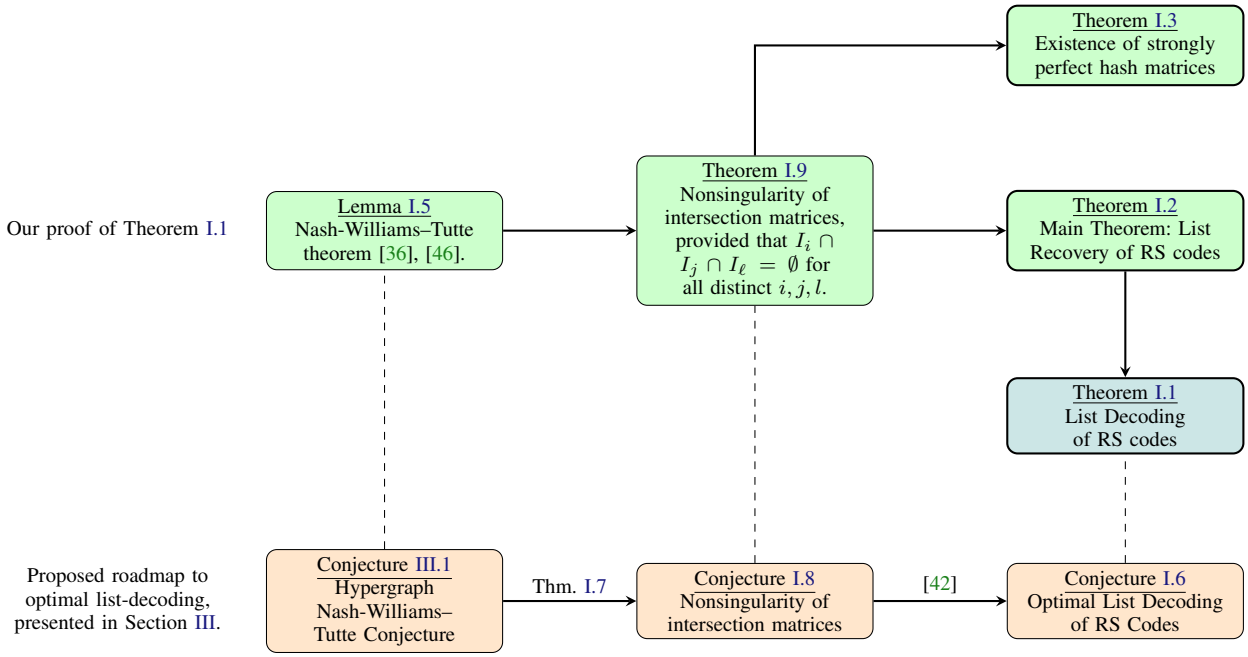


Figure 2. A diagram of the results and conjectures presented in this work. Solid arrows represent logical implications. Dashed lines indicate how the proposed roadmap to optimal list decoding parallels our proof of Theorem I.1.

### A. Cycle Spaces

We need the notion of the *cycle space* of a graph, which is typically defined over the boolean field $\mathbb{F}_2$ (see, e.g., [13]). Here we define it over an arbitrary field $\mathbb{F}$. An equivalent definition can be found in [5], where it is called the "circuit-subspace".

Let $K_t$ be the undirected complete graph with the vertex set $[t]$. Denote by $\{i,j\}$ the edge connecting vertices $i$ and $j$. Let $K_t^o$ be the oriented graph obtained by replacing $\{i,j\}$ with the directed edge $(i,j)$ for all $1 \le i < j \le t$. For a graph $G$ with vertex set $[t]$, an *oriented cycle* in $G$ is a set of directed edges of the form

$$C = \{(i_0,i_1),(i_1,i_2),\ldots,(i_{m-1},i_m)\}$$

where $m \ge 3$, $i_0,\ldots,i_{m-1}$ are distinct, $i_m = i_0$ and $\{i_{j-1},i_j\}$ is an edge of $G$ for all $j = 1,\ldots,m$.

Suppose $C$ is a union of edge-disjoint oriented cycles in $G$. Then $C$ is uniquely represented by a vector $u^C = (u^C_{\{i,j\}} : \{i,j\} \in \binom{[t]}{2}) \in \mathbb{F}^{\binom{t}{2}}$, defined for $1 \le i < j \le t$ by

$$u^C_{\{i,j\}} = \begin{cases} 1 & (i,j) \in C, \\ -1 & (j,i) \in C, \\ 0 & \text{else.} \end{cases}$$

Hence, the sign of a nonzero coordinate $u^C_{\{i,j\}}$ indicates whether the orientation of $\{i,j\}$ in $C$ complies with its orientation in $K_t^o$. We further assume that the coordinates of $u^C$ are ordered by the aforementioned lexicographic order on $\binom{[t]}{2}$.

Denote by $C(G) \subseteq \mathbb{F}^{\binom{t}{2}}$ the subspace spanned by the set of vectors

$$\{u^C : C \text{ is an oriented cycle in } G\}$$

over $\mathbb{F}$. We call $C(G)$ the *cycle space* of $G$ over $\mathbb{F}$. We are particularly interested in the cycle space $C(K_t)$ of $K_t$. For distinct $i,j,\ell \in [t]$, denote by $\Delta_{ij\ell}$ the oriented cycle $\{(i,j),(j,\ell),(\ell,i)\}$ and call it an *oriented triangle*. We have the following lemma, generalizing [13, Theorem 1.9.5].

**Lemma II.1.** *The vector space $C(K_t) \subseteq \mathbb{F}^{\binom{t}{2}}$ has dimension $\binom{t-1}{2}$, and the set*

$$\mathcal{B}_t = \{u^{\Delta_{ijt}} : 1 \le i < j \le t-1\}$$

*is a basis of $C(K_t)$.*

The basis $\mathcal{B}_t$ is also viewed as a $\binom{t-1}{2} \times \binom{t}{2}$ matrix over $\mathbb{F}$ whose columns are labeled by the edges $\{i,j\}$ of $K_t$, according to the lexicographic order defined above. Moreover, the rows of $\mathcal{B}_t$ represent $u^{\Delta_{ijt}}$ for $1 \le i < j \le t-1$, and are labeled by $\{i,j\} \in \binom{[t-1]}{2}$, also according to the lexicographic order. For example, $\mathcal{B}_3 = (1,-1,1)$ and

$$\mathcal{B}_4 = \begin{pmatrix} 1 & & -1 & 1 & \\ & 1 & -1 & & 1 \\ & & 1 & -1 & 1 \end{pmatrix},$$

where the 6 columns are labeled and ordered lexicographically by $\{1,2\} < \{1,3\} < \{2,3\} < \{1,4\} < \{2,4\} < \{3,4\}$. Observe for example that the $\pm 1$ entries in the

first row correspond to the oriented triangle $\Delta_{124} = \{(1,2),(2,4),(4,1)\}$, where we have $-1$ on the column labeled by the edge $\{1,4\}$, since the directed edge $(4,1)$ in $\Delta_{124}$ has the opposite orientation from the orientation of the edge in $K_t^o$.

We remark that the above definition of $\mathcal{B}_t$, is given with respect to the fixed orientation of the edges of $K_t^o$, as with the definition of $u^C$ for any oriented cycle $C$. One may define $\mathcal{B}_t$ with respect to other orientations of edges, which corresponds to changing the signs in some columns. These definitions are all equivalent and the analysis in this paper holds for any orientation up to change of signs.

Moreover, when the characteristic of $\mathbb{F}$ is two, we recover the definition of $\mathcal{B}_t$ in [42] using the fact that $1 = -1$. While working in the case $\mathrm{char}(\mathbb{F}) = 2$ has the advantage that there is no need to distinguish the signs, the theory holds more generally over any field.

### B. $t$-Wise Intersection Matrices

We proceed to define $t$-wise intersection matrices, but we begin with a few preliminary definitions. Given $n$ variables or field elements $x_1, \ldots, x_n$, define the $n \times k$ Vandermonde matrix

$$V_k(x_1, \ldots, x_n) = \begin{pmatrix} 1 & x_1 & \cdots & x_1^{k-1} \\ & & \ddots & \\ 1 & x_n & \cdots & x_n^{k-1} \end{pmatrix}.$$

When the $x_i$'s are understood from the context, for $I \subseteq [n]$, we use the abbreviation $V_k(I) := V_k(x_i : i \in I)$ to denote the restriction of $V_k(x_1, \ldots, x_n)$ to the rows with indices in $I$.

Let $\mathcal{I}_k$ denote the identity matrix of order $k$. Next, we give the definition of $t$-wise intersection matrices.

**Definition II.2** ($t$-wise intersection matrices)**.** *For a positive integer $k$ and $t \geq 3$ subsets $I_1, \ldots, I_t \subseteq [n]$, the $t$-wise intersection matrix $M_{k,(I_1,\ldots,I_t)}$ is the $(\binom{t-1}{2}k + \sum_{1 \leq i < j \leq t} |I_i \cap I_j|) \times \binom{t}{2}k$ variable matrix with entries in $\mathbb{F}[x_1, \ldots, x_n]$, defined as*

$$\left( \frac{\mathcal{B}_t \otimes \mathcal{I}_k}{\mathrm{diag}\Big(V_k(I_i \cap I_j) : \{i,j\} \in \binom{[t]}{2}\Big)} \right),$$

*where $\otimes$ is tensor product of matrices and*

- *$\mathcal{B}_t \otimes \mathcal{I}_k$ is a $\binom{t-1}{2}k \times \binom{t}{2}k$ matrix with entries in $\{0, \pm 1\}$,*
- *$\mathrm{diag}\big(V_k(I_i \cap I_j) : \{i,j\} \in \binom{[t]}{2}\big)$ is a block diagonal matrix with blocks $V_k(I_i \cap I_j)$, ordered by the lexicographic order on $\{i,j\} \in \binom{[t]}{2}$. Note that this matrix has order $(\sum_{1 \leq i < j \leq t} |I_i \cap I_j|) \times \binom{t}{2}k$. If $I_i \cap I_j = \emptyset$ for some $i, j$, then $V_k(I_i \cap I_j)$ is of order $0 \times k$ and the $\{i,j\} \in \binom{[t]}{2}$ block of $k$ columns is a $\sum_{1 \leq i < j \leq t} |I_i \cap I_j| \times k$ zero matrix.*

The reader is referred to the Appendix of [21] for an example of a 4-wise intersection matrix. We note that when $t = 2$, $\mathcal{B}_t$ is an empty matrix and $M_{k,(I_1,I_2)}$ is simply a Vandermonde matrix.

For a vector $\alpha \in \mathbb{F}^n$, the evaluation of $M_{k,(I_1,\ldots,I_t)}$ at the vector $\alpha$ is denoted by $M_{k,(I_1,\ldots,I_t)}(\alpha)$, where each variable $x_i$ is assigned the value $\alpha_i$. Given subsets $I_1, \ldots, I_t \subseteq [n]$, we call the variable matrix $M_{k,(I_1,\ldots,I_t)}$ *nonsingular* if it contains at least one $\binom{t}{2}k \times \binom{t}{2}k$ submatrix whose determinant is a nonzero polynomial in $\mathbb{F}[x_1, \ldots, x_n]$.

The paper [42] connects the nonsingularity of intersection matrices to the list-decodability of RS codes. We will also use this connection to prove our main result, Theorem I.2.

## III. Hypergraph Nash-Williams–Tutte Conjecture

Throughout, we use $t$ as the number of vertices in a (hyper)graph. This variable corresponds to the same $t$ used in $t$-wise intersection matrices. A (multi)graph $G$ is called *$k$-partition-connected* if every partition $\mathcal{P}$ of the vertex set has at least $k(|\mathcal{P}| - 1)$ edges crossing the partition. By the Nash-Williams–Tutte theorem, this is equivalent to the graph having $k$ edge-disjoint spanning trees. The parameter $k$ here is the same $k$ used as the dimension of the Reed–Solomon code and the same $k$ used for the Vandermonde matrix degrees in the intersection matrices.

We say a hypergraph $H$ is *$k$-weakly-partition-connected*[6] if, for every partition $\mathcal{P}$ of the vertices of $H$, we have

$$\sum_{e \in E(H)} (\mathcal{P}(e) - 1) \geq k(|\mathcal{P}| - 1),$$

where $\mathcal{P}(e)$ is the number of parts of $\mathcal{P}$ that $e$ intersects. For example, any $k$-partition-connected graph is $k$-weakly-partition-connected as a hypergraph. As another example, $k$ copies of a hyperedge covering all $t$ vertices of $H$ is also $k$-weakly partition-connected.

An *edge-labeled* graph is a graph $G$ where each edge is assigned a label from some set $E$. Let $H$ be a hypergraph. A *tree-assignment* of $H$ is an edge-labeled graph $G$ obtained by replacing each edge $e$ of $H$ with a tree $F_e$ of $|e| - 1$ edges on the vertices of $e$. Furthermore, each edge of the graph $F_e$ is labeled with $e$. The graph $G$ is thus the union of the graphs $F_e$ for $e \in H$.

A *$k$-tree-decomposition* of a graph on $k(t-1)$ edges is a partition of its edges into $k$ edge-disjoint spanning trees $T_0, \ldots, T_{k-1}$. We say *tree-decomposition* when $k$ is understood. In an edge-labeled graph $T$ with edge-labels from some set $E$, let $v^T \in \mathbb{N}^E$ be the vector counting the

[6]There is also a notion of "$k$-partition-connected" for hypergraphs which uses $\min\{\mathcal{P}(e) - 1, 1\}$ in the sum. In other words, a hypergraph is $k$-partition-connected if any partition $\mathcal{P}$ has at least $k(|\mathcal{P}|-1)$ crossing edges. This notion admits a Nash-Williams–Tutte type theorem: any $k$-partition-connected hypergraph can be decomposed into $k$ 1-partition-connected hypergraphs [18]

edge-labels in $T$. Specifically, $v_e^T$ is the number of edges of label $e$ in $T$. For a tree-decomposition $(T_0, \cdots, T_{k-1})$ of an edge-labeled graph, define its *signature* $v^{(T_0,\ldots,T_{k-1})}$ by

$$v^{(T_0,\ldots,T_{k-1})} := \sum_{i=0}^{k-1} i \cdot v^{T_i}.$$

An edge-labeled graph $G$ on $t$ vertices is called *$k$-distinguishable* if $G$ has $k(t-1)$ edges and there exists a tree-decomposition $T_0, \ldots, T_{k-1}$ of $G$ with a unique signature. That is, for any tree-decomposition $T'_0, \ldots, T'_{k-1}$ with the same signature $v^{(T'_0,\ldots,T'_{k-1})} = v^{(T_0,\ldots,T_{k-1})}$, we have $T'_i = T_i$ for $i = 0, \ldots, k-1$.

With these definitions, we can now conjecture a hypergraph version of the Nash-Williams–Tutte theorem.

**Conjecture III.1.** *Let $t$ and $k$ be positive integers. Every $k$-weakly-partition-connected hypergraph $H$ on $t$ vertices has a $k$-distinguishable tree-assignment.*

## ACKNOWLEDGMENT

Zeyu Guo was supported by NSF-BSF grant CCF-1814629 and 2017732 and the Milgrom family grant for Collaboration between the Technion and the University of Haifa. This work was done while he was at the University of Haifa. He wants to thank Noga Ron-Zewi for helpful discussions.

Ray Li is supported by NSF GRFP grant DGE-1656518 and by Jacob Fox's Packard Fellowship. He thanks Bruce Spang for helpful discussions.

Chong Shangguan is partially supported by the National Key Research and Development Program of China under Grant No. 2020YFA0712100, National Natural Science Foundation of China under Grant No. 12101364, and the Qilu Scholar Program of Shandong University. Part of the work was done while he was a postdoc at Tel Aviv University and was supported by the Israel Science Foundation (ISF grant number 1030/15).

Itzhak Tamo is partially supported by the European Research Council (ERC grant number 852953), and by the Israel Science Foundation (ISF grant number 1030/15).

Mary Wootters is supported by NSF grant CCF-1844628 and NSF-BSF grant CCF-1814629, and by a Sloan Research Fellowship. She thanks Noga Ron-Zewi for helpful discussions.

We thank Karthik Chandrasekaran for helpful discussions about hypergraph packing theorems and for the reference [10].

## REFERENCES

[1] N. Alon and S. Gutner, "Balanced families of perfect hash functions and their applications," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2007, pp. 435–446.

[2] ——, "Balanced hashing, color coding and approximate counting," in *International Workshop on Parameterized and Exact Computation*. Springer, 2009, pp. 1–16.

[3] N. Alon and M. Naor, "Derandomization, witnesses for Boolean matrix multiplication and construction of perfect hash functions," *Algorithmica*, vol. 16, no. 4-5, pp. 434–449, 1996.

[4] E. Ben-Sasson, S. Kopparty, and J. Radhakrishnan, "Subspace polynomials and limits to list decoding of Reed-Solomon codes," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 113–120, Jan 2010.

[5] N. Biggs, N. L. Biggs, and B. Norman, *Algebraic graph theory*. Cambridge University Press, 1993, vol. 67.

[6] S. R. Blackburn, "Perfect hash families: probabilistic methods and explicit constructions," *Journal of Combinatorial Theory, Series A*, vol. 92, no. 1, pp. 54–60, 2000.

[7] ——, "Combinatorial schemes for protecting digital content," in *Surveys in combinatorics, 2003 (Bangor)*, ser. London Math. Soc. Lecture Note Ser. Cambridge University Press, 2003, vol. 307, pp. 43–78.

[8] S. R. Blackburn and P. R. Wild, "Optimal linear perfect hash families," *Journal of Combinatorial Theory Series A*, vol. 83, no. 2, pp. 233–250, 1998.

[9] J.-Y. Cai, A. Pavan, and D. Sivakumar, "On the hardness of permanent," in *Annual Symposium on Theoretical Aspects of Computer Science*. Springer, 1999, pp. 90–99.

[10] G. Călinescu, C. Chekuri, and J. Vondrák, "Disjoint bases in a polymatroid," *Random Structures & Algorithms*, vol. 35, no. 4, pp. 418–430, 2009.

[11] Q. Cheng and D. Wan, "On the list and bounded distance decodability of Reed-Solomon codes," *SIAM J. Comput.*, vol. 37, no. 1, pp. 195–209, Apr. 2007. [Online]. Available: http://dx.doi.org/10.1137/S0097539705447335

[12] J. Cheriyan and M. R. Salavatipour, "Packing element-disjoint steiner trees," *ACM Transactions on Algorithms (TALG)*, vol. 3, no. 4, pp. 47–es, 2007.

[13] R. Diestel, *Graph theory*, 5th ed., ser. Graduate Texts in Mathematics. Springer, Berlin, 2017, vol. 173. [Online]. Available: https://doi.org/10.1007/978-3-662-53622-3

[14] D. Doron and M. Wootters, "High-probability list-recovery, and applications to heavy hitters." in *Electron. Colloquium Comput. Complex.*, vol. 27, 2020, p. 162.

[15] R. Dougherty, "Hash families and applications to t-restrictions," Ph.D. dissertation, Doctoral Dissertation Arizona State University, 2019.

[16] P. Elias, "List decoding for noisy channels," *Wescon Convention Record, Part 2, Institute of Radio Engineers*, pp. 99–104, 1957.

[17] A. Ferber, M. Kwan, and L. Sauermann, "List-decodability with large radius for reed-solomon codes," *arXiv preprint arXiv:2012.10584*, 2020.

[18] A. Frank, T. Király, and M. Kriesell, "On decomposing a hypergraph into k connected sub-hypergraphs," *Discrete Applied Mathematics*, vol. 131, no. 2, pp. 373–383, 2003.

[19] M. L. Fredman and J. Komlós, "On the size of separating systems and families of perfect hash functions," *SIAM Journal on Algebraic Discrete Methods*, vol. 5, no. 1, pp. 61–68, 1984.

[20] E. Goldberg, C. Shangguan, and I. Tamo, "List-decoding and list-recovery of reed-solomon codes beyond the Johnson radius for any rate," *arXiv preprint arXiv:2105.14754*, 2021.

[21] Z. Guo, R. Li, C. Shangguan, I. Tamo, and M. Wootters, "Improved list-decodability of reed–solomon codes via tree packings," *arXiv preprint arXiv:2011.04453*, 2020.

[22] V. Guruswami and A. Rudra, "Limits to list decoding Reed–Solomon codes," *IEEE Trans. Inform. Theory*, vol. 52, no. 8, pp. 3642–3649, Aug. 2006. [Online]. Available: https://doi.org/10.1109/TIT.2006.878164

[23] V. Guruswami and C. Wang, "Linear-algebraic list decoding for variants of Reed-Solomon codes," *IEEE Trans. Inform. Theory*, vol. 59, no. 6, pp. 3257–3268, June 2013.

[24] V. Guruswami and A. Rudra, "Explicit codes achieving list decoding capacity: Error-correction with optimal redundancy," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 135–150, 2008.

[25] V. Guruswami, A. Rudra, and M. Sudan, "Essential coding theory," *Draft available at http://cse.buffalo.edu/faculty/atri/courses/coding-theory/book/*, 2019.

[26] V. Guruswami and M. Sudan, "Improved decoding of Reed–Solomon and algebraic-geometry codes," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 1757–1767, 1999.

[27] ——, "Extensions to the johnson bound," *Manuscript, February*, 2001.

[28] V. Guruswami, C. Umans, and S. Vadhan, "Unbalanced expanders and randomness extractors from Parvaresh–Vardy codes," *Journal of the ACM (JACM)*, vol. 56, no. 4, pp. 1–34, 2009.

[29] S. Johnson, "A new upper bound for error-correcting codes," *IRE Transactions on Information Theory*, vol. 8, no. 3, pp. 203–207, 1962.

[30] S. Kopparty, N. Ron-Zewi, S. Saraf, and M. Wootters, "Improved decoding of folded Reed-Solomon and multiplicity codes," in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2018, pp. 212–223.

[31] S. Kopparty, "List-decoding multiplicity codes," *Theory of Computing*, vol. 11, no. 1, pp. 149–182, 2015.

[32] J. Körner, "Fredman–komlós bounds and information theory," *SIAM Journal on Algebraic Discrete Methods*, vol. 7, no. 4, pp. 560–570, 1986.

[33] J. Korner and K. Marton, "New bounds for perfect hashing via information theory," *European Journal of Combinatorics*, vol. 9, no. 6, pp. 523–530, 1988.

[34] B. Lund and A. Potukuchi, "On the list recoverability of randomly punctured codes," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2020)*, vol. 176, 2020, pp. 30:1–30:11.

[35] K. Mehlhorn, *Data structures and algorithms 1: Sorting and searching*, ser. EATCS Monographs on Theoretical Computer Science. Springer-Verlag, Berlin, 1984.

[36] C. S. J. A. Nash-Williams, "Edge-disjoint spanning trees of finite graphs," *Journal of the London Mathematical Society*, vol. 1, no. 1, pp. 445–450, 1961.

[37] I. Newman and A. Wigderson, "Lower bounds on formula size of Boolean functions using hypergraph entropy," *SIAM Journal on Discrete Mathematics*, vol. 8, no. 4, pp. 536–542, 1995.

[38] A. Nilli, "Perfect hashing and probability," *Combinatorics, Probability & Computing*, vol. 3, no. 3, pp. 407–409, 1994.

[39] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, 1960. [Online]. Available: https://doi.org/10.1137/0108018

[40] A. Rudra and M. Wootters, "Every list-decodable code for high noise has abundant near-optimal rate puncturings," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, ser. STOC 2014, 2014, pp. 764–773. [Online]. Available: http://doi.acm.org/10.1145/2591796.2591797

[41] C. Shangguan and G. Ge, "Separating hash families: A Johnson-type bound and new constructions," *SIAM Journal on Discrete Mathematics*, vol. 30, no. 4, pp. 2243–2264, 2016.

[42] C. Shangguan and I. Tamo, "Combinatorial list-decoding of Reed-Solomon codes beyond the Johnson radius," in *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing*, ser. STOC 2020, 2020, pp. 538–551.

[43] ——, "Degenerate turán densities of sparse hypergraphs," *Journal of Combinatorial Theory, Series A*, vol. 173, p. 105228, 2020.

[44] R. Singleton, "Maximum distance $q$-nary codes," *IEEE Trans. Inform. Theory*, vol. 10, no. 2, pp. 116–118, April 1964.

[45] M. Sudan, L. Trevisan, and S. Vadhan, "Pseudorandom generators without the xor lemma," *Journal of Computer and System Sciences*, vol. 62, no. 2, pp. 236–266, 2001.

[46] W. T. Tutte, "On the problem of decomposing a graph into n connected factors," *Journal of the London Mathematical Society*, vol. 1, no. 1, pp. 221–230, 1961.

[47] J. M. Wozencraft, "List decoding," *Quarterly Progress Report, Research Laboratory of Electronics, MIT*, vol. 48, pp. 90–95, 1958.

[48] C. Xing and C. Yuan, "Beating the probabilistic lower bound on perfect hashing," *arXiv preprint arXiv:1908.08792*, 2019.