Two-Source Dispersers for Polylogarithmic Entropy and Improved Ramsey Graphs

Gil Cohen
Weizmann Institute of Science
234 Herzl Street
Rehovot, Israel
coheng@gmail.com

ABSTRACT

In his influential 1947 paper that inaugurated the probabilistic method, Erdős proved the existence of $2\log n$-Ramsey graphs on n vertices. Matching Erdős' result with a constructive proof is considered a central problem in combinatorics, that has gained a significant attention in the literature. The state of the art result was obtained in the celebrated paper by Barak, Rao, Shaltiel, and Wigderson who constructed a $2^{2^{(\log\log n)^{1-\alpha}}}$-Ramsey graph, for some small universal constant $\alpha > 0$.

In this work, we significantly improve this result and construct $2^{(\log\log n)^c}$-Ramsey graphs, for some universal constant c. In the language of theoretical computer science, this resolves the problem of explicitly constructing dispersers for two n-bit sources with entropy polylog(n). In fact, our disperser is a zero-error disperser that outputs a constant fraction of the entropy. Prior to this work, such dispersers could only support entropy $\Omega(n)$.

Categories and Subject Descriptors

F.0 [**Theory of Computation**]: General

General Terms

Theory

Keywords

Ramsey graphs, explicit constructions, zero-error dispersers

1. INTRODUCTION

Ramsey theory is a branch of combinatorics that studies the unavoidable presence of local structure in globally unstructured objects. In the paper that pioneered this field of study, Ramsey [Ram28] considered an instantiation of this phenomena in graph theory.

Definition 1. A graph on n vertices is called k-Ramsey if it contains no clique or independent set of size k.

STOC'16, June 19–21, 2016, Cambridge, MA, USA
© 2016 ACM. 978-1-4503-4132-5/16/06...$15.00
http://dx.doi.org/10.1145/2897518.2897530

Ramsey showed that there does not exist a graph on n vertices that is $\log(n)/2$-Ramsey. In his influential paper that inaugurated the probabilistic method, Erdős [Erd47] complemented Ramsey's result and showed that most graphs on n vertices are $2\log n$-Ramsey. Unfortunately, Erdős' argument is non-constructive and one does not obtain from Erdős' proof an example of a graph that is $2\log n$-Ramsey. Erdős offered a $100 prize for matching his result, up to any multiplicative constant factor, by a constructive proof. That is, coming up with an explicit construction of an $O(\log n)$-Ramsey graph. Erdős' challenge gained a significant attention in the literature [Abb72, Nag75, Fra77, Chu81, FW81, Nao92, Alo98, Gro01, PR04, Bar06, BKS+10, BRSW12]. Other works studied the difficulty of constructing Ramsey graphs [Gop14] and suggested routes towards constructing improved Ramsey graphs [GKRTS05].

The notion of explicitness was formalized in the computational era. While, classically, a succinct mathematical formula was widely considered to be an explicit description, complexity theory suggests a more relaxed, and arguably more natural interpretation of explicitness. An object is deemed explicit if one can efficiently construct that object from scratch. More specifically, a graph on n vertices is explicit if given the labels of any two vertices u, v, one can efficiently determine whether there is an edge connecting u, v in the graph. Since the description length of u, v is $2\log n$ bits, quantitatively, by efficient we require that the running-time is polylog(n).

Ramsey graphs have an analogous definition for bipartite graphs. A bipartite graph on two sets of n vertices is bipartite k-Ramsey if it has no $k \times k$ complete or empty bipartite subgraph. One can show that a bipartite Ramsey graph induces a Ramsey graph with comparable parameters. Thus, constructing bipartite Ramsey-graphs is at least as hard as constructing Ramsey graphs, and it was believed to be a strictly harder problem. Nevertheless, the best known construction of Ramsey graphs is in fact bipartite. Furthermore, Erdős' argument holds as is for bipartite graphs.

In their celebrated paper, Barak *et al.* [BRSW12] gave an explicit construction of a bipartite $k(n)$-Ramsey graph on n vertices with $k(n) = 2^{2^{(\log\log n)^{1-\alpha}}}$, where $\alpha > 0$ is some small universal constant. In particular, $k(n) = 2^{o(\log n)}$ is sub-exponential in the desired value, namely, in $2\log n$. In this paper we give an explicit construction of a bipartite $k(n)$-Ramsey graph with $k(n)$ that is pseudo-polynomial in the desired value. [1]

[1]A function $f\colon \mathbb{N} \to \mathbb{N}$ is *pseudo-polynomial* if there exist

THEOREM 1. *There exists an explicit bipartite* $2^{(\log\log n)^c}$*-Ramsey graph on* n *vertices, where* c *is some universal constant.*

1.1 Two-Source Zero-Error Dispersers

In the language of theoretical computer science, Theorem 1 translates to a disperser for two independent n-bit sources with entropy $O(\log^c n)$. We first recall some basic definitions.

Definition 2. The *min-entropy* of a random variable X is defined by $H_\infty(X) = \min_{x\in\sup(X)} -\log_2(\mathbf{Pr}[X=x])$. If X is supported on $\{0,1\}^n$, we define the *min-entropy rate* of X by $H_\infty(X)/n$. In such case, if X has min-entropy k or more, we say that X is an (n,k)-source.

Definition 3. A function $\mathsf{Disp}\colon \{0,1\}^n\times\{0,1\}^n \to \{0,1\}^m$ is called a *two-source zero-error disperser* for entropy k if for any two independent (n,k)-sources X,Y, it holds that $\sup(\mathsf{Disp}(X,Y)) = \{0,1\}^m$.

Note that a two-source zero-error disperser for entropy k, with a single output bit, is equivalent to a bipartite 2^k-Ramsey graph on 2^n vertices on each side. Constructing two-source dispersers for polylogarithmic entropy is considered a central problem in pseudorandomness, that we resolve in this paper. Indeed, a $2^{\mathrm{poly}(\log\log n)}$-Ramsey graph on n vertices is equivalent to a disperser for entropy $\mathrm{polylog}(n)$. From the point of view of dispersers, it is easier to see how challenging is Erdős' goal of constructing $O(\log n)$-Ramsey graphs. Indeed, these are equivalent to dispersers for entropy $\log(n)+O(1)$. Even a disperser for entropy $O(\log n)$ does not meet Erdős' goal as it translates to a $\mathrm{polylog}(n)$-Ramsey graph.

While Theorem 1 already yields a two-source zero-error disperser for polylogarithmic entropy, it is desired to construct dispersers with many output bits. Our construction has this property.

THEOREM 2. *There exists an explicit two-source zero-error disperser for* n*-bit sources having entropy* $k = \mathrm{polylog}(n)$*, with* $m = \Omega(k)$ *output bits.*

Theorem 2 gives an explicit zero-error disperser for polylogarithmic entropy with many output bits. Prior to this work, the state of the art zero-error disperser with a super constant number of output bits, due to Gabizon and Shaltiel [GS08], required entropy $k = \Omega(n)$. In fact, partially motivated by applications to data structures [FN93], in [GS08] a stronger variant of a two-source zero-error disperser was constructed, in which every element in the range is obtained with probability at least $\delta = \delta(n)$. Our construction has this property as well.

1.2 Implicit $O(1)$ Probe Search

In this section we describe an application of our construction to data structures. In the probe search problem one wants to store a set $S \subseteq \{0,1\}^n$ of size $|S| = 2^k$ in an ordered table T of the same size, where every entry in T contains exactly one element of S. We say that T supports q queries if given $x \in \{0,1\}^n$ one can determine whether $x \in S$ by probing q entries of T. Note that the only freedom one has when designing T is the order in which the elements of S are placed.

constants c, m_0 such that $f(m) \le 2^{(\log m)^c}$ for all $m > m_0$.

Regardless of the value of n, one can always store S in a table T according to a predetermined order of $\{0,1\}^n$ so to support $q = k$ queries. Following Yao [Yao81] and Fiat and Naor [FN93], the regime of parameters we consider only allows for constant q, independent of n, k. In [Yao81, FN93] it was shown that for supporting constant q, k must be a fast enough growing function of n. On the positive side, Fiat and Naor [FN93] showed how to construct tables supporting constant q for $k = \delta n$, for any constant $\delta > 0$. This result was later improved by Gabizon and Shaltiel [GS08] to $k = n^\delta$ for any constant $\delta > 0$, while maintaining constant query complexity.

Fiat and Naor [FN93] reduced the probe search problem to the design of a combinatorial object they called a "rainbow". This object can be constructed given a certain kind of dispersers. In particular, the output length of the disperser must be large. Although we do not delve to the details in this proceeding version, we remark that our dispersers are sufficient for this reduction to go through, and by building on [FN93, GS08] we improve previous results to $k = \mathrm{polylog}(n)$ while maintaining constant query complexity.

We remark that, by inspection, one can show that this improvement also follows by using a recent construction of multi-source extractors [Li15b]. Though, as these extractors require more than two sources, the obtained query complexity is slightly larger.

1.3 Subsequent Work

In an exciting subsequent work, Chattopadhyay and Zuckerman [CZ15] gave a construction of a two-source extractor for $\mathrm{polylog}(n)$-entropy based on a very different set of ideas than ours. The error of their extractor is $n^{-\Omega(1)}$ and the number of output bits is 1. The latter was improved soon after by Li [Li15a] using similar techniques, with the same error parameter. As extractors with one output bit yields Ramsey graphs, the work of [CZ15] gives a second, very different, construction of Ramsey graphs matching our parameters.

2. OVERVIEW OF THE CONSTRUCTION AND ANALYSIS

In this section we present our disperser and give a fairly comprehensive overview of the proof, though we allow ourselves to be somewhat imprecise whenever this makes the exposition clearer. The formal proof can be found in the full version of this paper. We start by recalling the definition of a subsource. For ease of presentation we consider a somewhat relaxed definition. From this point on an (n,k)-source will refer to a random variable X that is uniformly distributed over some subset $S_X \subset \{0,1\}^n$ with size at least 2^k. A lemma by Chor and Goldreich [CG88] implies that, essentially, this assumption can be made without loss of generality. We say that a source X' is a subsource of X, and write $X' \subset X$, if $S_{X'} \subseteq S_X$. Further, the deficiency of X' in X is defined by $\log(|S_{X'}|/|S_X|)$.

Another notion we need is that of a block-source. For an even integer n, given an n-bit string x, we denote by $\mathsf{left}(x)$ and by $\mathsf{right}(x)$ the $n/2$-bit prefix and $n/2$-bit suffix of x, respectively. A random variable X is called a k-block-source if $\mathsf{left}(X)$ has min-entropy k. Furthermore, conditioned on any fixing of $\mathsf{left}(X)$ it holds that $\mathsf{right}(X)$ has min-entropy k. Following a long line of research, Li [Li15b] gave an explicit construction of an extractor for a block-source and a source

with polylogarithmic entropy. More precisely, Li designed an efficiently computable function $\mathsf{BExt}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ with the following property. For any k-block-source X and an independent (n,k)-source Y, with $k \geq \log^{12} n$, $\mathsf{BExt}(X,Y) \approx_\varepsilon U_m$, where $m = 0.9k$ and $\varepsilon = 2^{-k^{\Omega(1)}}$.

2.1 Entropy-Trees and Tree-Structured Sources

For the purpose of constructing a disperser it suffices to construct a disperser for subsources of the original sources. In this section we show that any source has a low-deficiency subsource that has a "nice" structure. Thus, it suffices to construct our disperser only for these nice sources.

Entropy-trees

An entropy-tree is a complete rooted binary tree T, where some of its nodes are labeled by one of the following labels: $\mathsf{H}, \mathsf{B}, \mathsf{F}$, which stand for high entropy, block-source, and fixed, respectively. The nodes of an entropy-tree are labeled according to rules that capture any possible entropy structure of a source. The rules are:

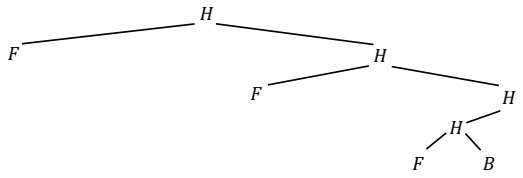

Figure 1: An example of an entropy-tree. Unlabeled nodes and edges to them are omitted.

- The root of T, denoted by $\mathsf{root}(T)$, is labeled by either H or B.
- There is exactly one node in T, denoted by $v_{\mathsf{B}}(T)$, that is labeled by B.
- If v is a non-leaf that has no label, or otherwise labeled by F or B, then its sons have no label.
- If v is a non-leaf that is labeled by H then the sons of v can only be labeled according to the following rules:
  - If $\mathsf{leftSon}(v)$ is labeled by F then $\mathsf{rightSon}(v)$ is labeled by either H or B.
  - If $\mathsf{leftSon}(v)$ is labeled by H or B then $\mathsf{rightSon}(v)$ has no label.

With every entropy-tree T, we associate a path that we call the *entropy-path* of T. This is the unique path from $\mathsf{root}(T)$ to $v_{\mathsf{B}}(T)$. We say that a path in T contains the entropy-path if it starts at $\mathsf{root}(T)$ and goes through $v_{\mathsf{B}}(T)$. Note that we allow an entropy-tree to have nodes that are descendants of $v_{\mathsf{B}}(T)$. We just do not allow these nodes to have labels.

Tree-structured sources

Now that we have defined entropy-trees, we can say what does it mean for a source to have a T-structure, for some entropy-tree T. To this end we need to introduce some notations. Let n be an integer that is a power of 2. With a string $x \in \{0,1\}^n$, we associate a depth $\log n$ complete rooted binary tree, where with each node v of T we associate a substring x_v of x in the following natural way: $x_{\mathsf{root}(T)} = x$; and for $v \neq \mathsf{root}(T)$, if v is the left son of its parent, then $x_v = \mathsf{left}(x_{\mathsf{parent}(v)})$; otherwise, $x_v = \mathsf{right}(x_{\mathsf{parent}(v)})$.

Let T be a depth $\log n$ entropy-tree. An n-bit source X is said to have a T-structure with parameter k if for any node v in T the following holds. If v is labeled by F then X_v is fixed; If v is labeled by H then $H_\infty(X_v) \geq k$; Otherwise, if v is labeled by B then X_v is a $\sqrt{k}$-block-source. One can prove that any (n,k)-source X, with $k = \Omega(\log^2 n)$, has a deficiency $2\log n$ subsource that has a tree-structure with parameter $\Omega(k)$. Thus, for the purpose of constructing dispersers, we may assume that we are given two independent samples from tree-structured sources rather than from general sources.

The challenge-response mechanism

The challenge-response mechanism was introduced in [BKS+10] and was further developed by [BRSW12]. Roughly speaking, this is a mechanism that allows one to distinguish between the case that a random variable is fixed from the case that it has a sufficient amount of entropy. We now present the mechanism.

For integers $\ell < n$, the challenge-response mechanism is a poly(n)-time computable function $\mathsf{Resp}\colon \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^\ell \to \{\mathsf{fixed}, \mathsf{hasEntropy}\}$ with the following property. For any two independent $(n, \text{polylog}(n))$-sources X, Y, and for any function $\mathsf{Ch}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^\ell$, the following holds:

- If $\mathsf{Ch}(X,Y)$ is fixed to a constant then there exist deficiency ℓ subsources $X' \subset X$, $Y' \subset Y$, such that $\mathbf{Pr}_{(x,y)\sim(X',Y')}[\mathsf{Resp}(x,y,\mathsf{Ch}(x,y)) = \mathsf{fixed}] = 1$.
- If for any deficiency ℓ subsources $\hat{X} \subset X$, $\hat{Y} \subset Y$ it holds that $H_\infty(\mathsf{Ch}(\hat{X},\hat{Y})) \geq k$, then

$$\mathbf{Pr}_{(x,y)\sim(X,Y)}[\mathsf{Resp}(x,y,\mathsf{Ch}(x,y)) = \mathsf{hasEntropy}] \geq 1 - 2^{-k}.$$

2.2 Identifying the Entropy-Path

Tree-structured sources certainly seem nicer to work with than general sources. However, it is still not clear what good is this structure for if we do not have any information regarding the entropy-tree, and in particular regarding the entropy-path. Remarkably, Barak *et al.* [BRSW12] were able to identify the entropy-path of the entropy-tree T given just one sample from $x \sim X$, where X is a T-structured source, and one sample from $y \sim Y$, where Y is a general source that is independent of X. We now turn to describe the algorithm used by [BRSW12]. Before doing so, we remark that Barak *et al.* proved something somewhat different. Indeed, they considered a variant of entropy-trees and had to prove something somewhat different than what we need. In particular, their algorithm did not identify the entropy-path per se. Nevertheless, their proof can be adapted in a straightforward manner to obtain the result we need.

What does it mean to identify the entropy-path?

What do we mean by saying that an algorithm identifies the entropy-path of an entropy-tree T? This is an algorithm that on input $x, y \in \{0,1\}^n$, outputs a depth $\log n$ rooted complete binary tree and a marked root-to-leaf path on that tree, denoted by $p_{\mathsf{obs}}(x,y)$ – the *observed* entropy-path. Ideally, the guarantee of the algorithm would have been

the following: If x is sampled from a T-structured source X and y is sampled independently from a source Y, then $p_{\mathsf{obs}}(x,y)$ contains the entropy-path of T with probability 1 over $(x,y) \sim (X,Y)$. That is, for any $(x,y) \in \sup((X,Y))$, if we draw the computed path $p_{\mathsf{obs}}(x,y)$ on the entropy-tree T then this path starts at $\mathsf{root}(T)$ and goes through $v_{\mathsf{B}}(T)$.

Note that the path $p_{\mathsf{obs}}(x,y)$ is allowed to continue arbitrarily after visiting $v_{\mathsf{B}}(T)$. Asking that $p_{\mathsf{obs}}(x,y)$ will stop exactly at $v_{\mathsf{B}}(T)$ is a very strong requirement. In particular, it will conclude the construction of the disperser. Indeed, once the block-source $X_{v_{\mathsf{B}}(T)}$ is found, one can simply output $\mathsf{BExt}(X_{v_{\mathsf{B}}(T)}, Y)$.

This was an ideal version of what we mean by identifying an entropy-path. For our needs, we will be satisfied with a weaker guarantee. Following [BRSW12], we will show that there exist low-deficiency subsources $X' \subset X$, $Y' \subset Y$, such that with high probability over $(x,y) \sim (X',Y')$ it holds that $p_{\mathsf{obs}}(x,y)$ contains the entropy-path of T.

The fact that we only have a guarantee on low-deficiency subsources is good enough for us as we are aiming for a disperser. The fact that there is an error (that did not appear in the analysis of [BRSW12]) should be handled with some care. Indeed, note that by moving to a deficiency d subsource, an ε error in the original source can grow to at most $2^d \cdot \varepsilon$ restricted to the subsource. We will make sure that the error is negligible compared to the deficiency we consider in the rest of the analysis. Thus, from here on we will forget about the error introduced in this step of identifying the entropy-path.

The algorithm of [BRSW12] for identifying the entropy-path

We now describe the algorithm used by [BRSW12] for identifying the entropy-path of an entropy-tree T. Note that if $\mathsf{root}(T) = v_{\mathsf{B}}(T)$ then any observed entropy-path will contain $v_{\mathsf{B}}(T)$. So, we may assume that this is not the case. Let v be the parent of $v_{\mathsf{B}}(T)$ in T. As a first step, we want to determine which of the two sons of v is $v_{\mathsf{B}}(T)$. To this end, node v declares that its left son is $v_{\mathsf{B}}(T)$ if and only if

$$\mathsf{Resp}\left(x_v, y, \mathsf{BExt}\left(x_{\mathsf{leftSon}(v)}, y\right)\right) = \mathsf{hasEntropy}. \qquad (1)$$

Lets pause for a moment to introduce some notations. If Equation (1) holds, we say that node v (x,y)-*favors its left-son*; otherwise, we say that v (x,y)-*favors its right son*. Moreover, we define the *good son* of v to be $v_{\mathsf{B}}(T)$. More generally, for a node $u \neq v_{\mathsf{B}}(T)$ that is an ancestor of $v_{\mathsf{B}}(T)$, we define the *good son* of u to be its unique son that is an ancestor of $v_{\mathsf{B}}(T)$. Note that by following the good sons from $\mathsf{root}(T)$ to $v_{\mathsf{B}}(T)$ one recovers the entropy-path of T. Thus, one correctly identifies the entropy-path of T on input x,y if and only if any ancestor of $v_{\mathsf{B}}(T)$ on the entropy-path of T (x,y)-favors its good son.

One can show that if $X_{\mathsf{leftSon}(v)}$ is fixed then Equation (1) holds with probability 0 on some low-deficiency subsources of X,Y. Further, by the challenge-response mechanism, one can show that if $\mathsf{leftSon}(v) = v_{\mathsf{B}}(T)$ then with high probability over (X,Y), Equation (1) holds. Observe that by the definition of an entropy-tree, these are the only two possible cases.

We showed how $v_{\mathsf{B}}(T)$ can convince its parent v that it is its good son. The trick was to use the block-source-ness of $X_{v_{\mathsf{B}}(T)}$ so to generate a proper challenge. Considering one step further, we ask the following: If u is the parent of v, how can v convince u that it is its good son? After all, v is not a block-source. The elegant solution of Barak *et al.* is as follows. Given $x,y \in \{0,1\}^n$, the challenge of v will contain not only $\mathsf{BExt}(x_v, y)$ but also $\mathsf{BExt}(x_w, y)$, where w is v's (x,y)-favored son. Thus, if v's favored son happens to be its good son $v_{\mathsf{B}}(T)$, the challenge posed by v will not be responded by u. More generally, a node v decides which of its two sons it (x,y)-favors not according to Equation (1) but rather according to whether or not

$$\mathsf{Resp}\left(x_v, y, \mathsf{GSC}\left(x_{\mathsf{leftSon}(v)}, y\right)\right) = \mathsf{hasEntropy}, \qquad (2)$$

where $\mathsf{GSC}(x_{\mathsf{leftSon}(v)}, y)$ is a matrix with at most $\log n$ rows (as the depth of the tree) that contains $\mathsf{BExt}(x_{\mathsf{leftSon}(v)}, y)$ as a row, as well as $\mathsf{BExt}(x_w, y)$, where w is the (x,y)-favored son of $\mathsf{leftSon}(v)$, and also $\mathsf{BExt}(x_r, y)$, where r is the (x,y)-favored son of w, etc.

2.3 The Strategy For The Rest of Our Construction

To carry the analysis of our disperser, we require even more structure from our sources than the structure required by [BRSW12]. First, we require *both* X and Y to have a tree-structure. In previous works [BKS+10, BRSW12], the second source Y was used mainly to "locate the entropy" of the source X, and the only assumption on Y was that it has a sufficient amount of entropy for this purpose. We, however, will make use of the structure of Y as well.

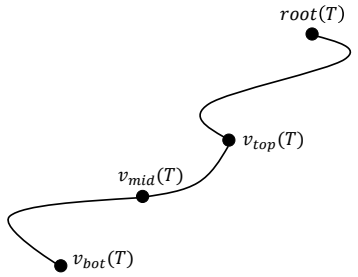

Figure 2: The "triple block-source" structure of an entropy-tree.

Second, we need both X and Y to have a "triple block-source" structure. That is, we assume that X has a T_X-structure with a node $v_{\mathsf{top}}(T_X)$ corresponding to the block-source $X_{v_{\mathsf{top}}(T_X)}$. We then assume that $\mathsf{left}(X_{v_{\mathsf{top}}(T_X)})$ has its own tree-structure with a node $v_{\mathsf{mid}}(T_X)$ corresponding to a second block-source $X_{v_{\mathsf{mid}}(T_X)}$ lying inside $\mathsf{left}(X_{v_{\mathsf{top}}(T_X)})$. Finally, we require that $\mathsf{left}(X_{v_{\mathsf{mid}}(T_X)})$ has its own tree-structure with a node $v_{\mathsf{bot}}(T_X)$ that corresponds to a third block-source $X_{v_{\mathsf{bot}}(T_X)}$ that lies inside $\mathsf{left}(X_{v_{\mathsf{mid}}(T_X)})$. The same goes for Y. Namely, Y also has a triple block-source structure. In particular, the entropy-tree of Y, denoted by T_Y, has nodes that we denote by $u_{\mathsf{top}}(T_Y)$, $u_{\mathsf{mid}}(T_Y)$, and $u_{\mathsf{bot}}(T_Y)$, analogous to $v_{\mathsf{top}}(T_X)$, $v_{\mathsf{mid}}(T_X)$, and $v_{\mathsf{bot}}(T_X)$ in T_X. We allow ourselves to change the definition of an entropy-tree given in the previous section so that it will capture this "triple block-source" structure, but the reader should not worry about these details at this point.

Given this structure of the sources, we are ready to give a high-level overview of our construction. In the subsequent sections of the overview (Section 2.4 and Section 2.5), we give

further details. Let X be a T_X-structured source and let Y be a T_Y-structured source, for some entropy-trees T_X, T_Y. At the first step, the disperser identifies the entropy-path of T_X and the entropy-path of T_Y using the algorithm of [BRSW12]. More precisely, given the samples $x \sim X$,$y \sim Y$, we compute two paths denoted by $p_{\mathsf{obs}}(x,y) = v_0(x,y), \ldots, v_{\log(n)-1}(x,y)$, $q_{\mathsf{obs}}(x,y) = u_0(x,y), u_1(x,y), \ldots, u_{\log(n)-1}(x,y)$. This step must be done with some care. From technical reasons (related to the way the error term behaves when moving to subsources), we cannot use x, y to first find the entropy-path of T_X and then to find the entropy-path of T_Y. Thus, in some sense, the two paths must be computed simultaneously.

At this point, ignoring some small error term, we have that there exist low-deficiency subsources $X' \subset X$, $Y' \subset Y$, such that for any $(x,y) \in \sup((X',Y'))$ it holds that $p_{\mathsf{obs}}(x,y)$ (resp. $q_{\mathsf{obs}}(x,y)$) contains the entropy-path of T_X (resp. T_Y). In particular, we have that $v_{\mathsf{depth}(v_{\mathsf{top}}(T_X))}(X',Y')$ is fixed to $v_{\mathsf{top}}(T_X)$, and the same holds for $v_{\mathsf{mid}}(T_X)$, $v_{\mathsf{bot}}(T_X)$, as well as for $u_{\mathsf{top}}(T_Y)$, $u_{\mathsf{mid}}(T_Y)$, and $u_{\mathsf{bot}}(T_Y)$. To keep the notations clean, we write X, Y for X', Y' in this proof overview.

At the second step of the algorithm, we identify $v_{\mathsf{mid}}(T_X)$ with high probability over subsources $X' \subset X$, $Y' \subset Y$. This sounds fantastic – having found $v_{\mathsf{mid}}(T_X)$, we can simply output $\mathsf{BExt}(X'_{v_{\mathsf{mid}}(T_X)}, Y')$ which is close to uniform. Unfortunately, however, the only way we know how to find $v_{\mathsf{mid}}(T_X)$ requires us to fix $\mathsf{left}(X'_{v_{\mathsf{mid}}(T_X)})$. That is, once found, $X'_{v_{\mathsf{mid}}(T_X)}$ is no longer a block-source. We elaborate on how to find $v_{\mathsf{mid}}(T_X)$ in Section 2.4. Then in Section 2.5, we show how to determine the output even after loosing the block-structure of $X_{v_{\mathsf{mid}}(T_X)}$.

2.4 Finding $v_{\mathsf{mid}}(T_X)$

Given $x, y \in \{0,1\}^n$, the key idea we use for identifying $v_{\mathsf{mid}}(T_X)$ on $p_{\mathsf{obs}}(x,y)$ lies in the design of a challenge that we call the *node-path challenge*.

The node-path challenge and $v^{\mathsf{obs}}_{\mathsf{mid}}(x,y)$

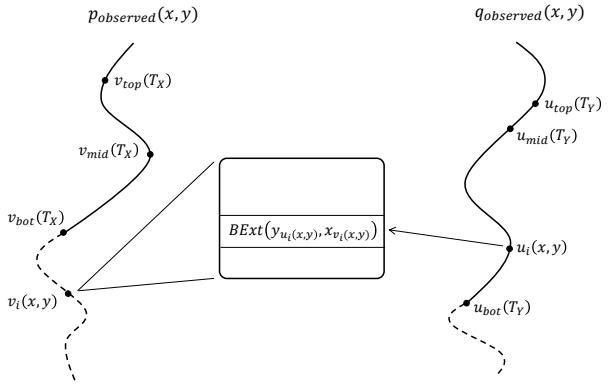

Figure 3: The node-path challenge.

Let v be a node in T_X, and let $q = w_0, \ldots, w_{\log(n)-1}$ be a root-to-leaf path in T_Y. We define the challenge $\mathsf{NPC}(x_v, y_q)$ to be the $\log(n)$-rows Boolean matrix such that for $i = 0, 1, \ldots, \log(n)-1$, $\mathsf{NPC}(x_v, y_q)_i = \mathsf{BExt}(y_{w_i}, x_v)$. We define $v^{\mathsf{obs}}_{\mathsf{mid}}(x,y)$ to be the node v on $p_{\mathsf{obs}}(x,y)$ with the largest depth such that

$$\mathsf{Resp}\left(x, y, \mathsf{NPC}\left(x_v, y_{q_{\mathsf{obs}}(x,y)}\right)\right) = \mathsf{hasEntropy}. \quad (3)$$

Informally speaking, based on the node-path challenge, a node on $p_{\mathsf{obs}}(x,y)$ uses the path $q_{\mathsf{obs}}(x,y)$ so to prove that it is $v_{\mathsf{mid}}(T_X)$.

Ideally, we would want to prove that $v^{\mathsf{obs}}_{\mathsf{mid}}(x,y) = v_{\mathsf{mid}}(T_X)$ for any $(x,y) \in \sup((X,Y))$. By now we know that this is too much to ask for, and in any case, it suffices to prove that there exist low-deficiency subsources $X' \subset X$, $Y' \subset Y$ such that with high probability over $(x,y) \sim (X',Y')$ it holds that $v^{\mathsf{obs}}_{\mathsf{mid}}(x,y) = v_{\mathsf{mid}}(T_X)$. Unfortunately, we will not be able to prove that either. What we will be able to show is that there exist strings α, β such that the following holds. Define $X_\alpha = X \mid \left(X_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))} = \alpha\right)$, $Y_\beta = Y \mid \left(Y_{\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))} = \beta\right)$, and let $i_{\mathsf{mid}}(T_X)$ denote the depth of $v_{\mathsf{mid}}(T_X)$.

The way we choose α, β is with respect to the error that we constantly ignore throughout this overview. Thus, assume that α, β are chosen in such a way that allows us to continue ignoring the error (this is done by a simple averaging argument). No further requirement is posed on α, β.

Proposition 1. There exist low-deficiency subsources $X_{\alpha,\beta} \subset X_\alpha$, $Y_{\alpha,\beta} \subset Y_\beta$, such that with high probability over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$, it holds that

$$\forall i > i_{\mathsf{mid}}(T_X) \quad \mathsf{Resp}\left(x, y, \mathsf{NPC}\left(x_{v_i(x,y)}, y_{q_{\mathsf{obs}}(x,y)}\right)\right) = \mathsf{fixed},$$

$$\mathsf{Resp}\left(x, y, \mathsf{NPC}\left(x_{v_{i_{\mathsf{mid}}(T_X)}(x,y)}, y_{q_{\mathsf{obs}}(x,y)}\right)\right) = \mathsf{hasEntropy}.$$

Note that by the way we defined $v^{\mathsf{obs}}_{\mathsf{mid}}(x,y)$, Proposition 1 yields that $v^{\mathsf{obs}}_{\mathsf{mid}}(x,y) = v_{\mathsf{mid}}(T_X)$ with high probability over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$. In particular, this gives us an algorithm for computing $v_{\mathsf{mid}}(T_X)$ – simply go up the computed path $p_{\mathsf{obs}}(x,y)$ until a node v is found for which Equation (3) holds. In the rest of this section we prove Proposition 1.

Proposition 1 has two parts. First, it states that the node-path challenges associated with nodes below $v_{i_{\mathsf{mid}}(T_X)}(x,y)$ on the path $p_{\mathsf{obs}}(x,y)$ are responded with high probability over x, y that are sampled from some low-deficiency subsources of X_α, Y_β. Second, the node-path challenge associated with $v_{i_{\mathsf{mid}}(T_X)}(x,y)$ is left unresponded with high probability over the samples.

Recall that, ignoring a small error term, we assume that $v_{i_{\mathsf{mid}}(T_X)}(x,y) = v_{\mathsf{mid}}(T_X)$. Lets first consider the nodes below $v_{\mathsf{mid}}(T_X)$ on $p_{\mathsf{obs}}(x,y)$. Naturally, we want to use the challenge-response mechanism. For that we must find low-deficiency subsources $X'_\alpha \subset X_\alpha$, $Y'_\beta \subset Y_\beta$ such that for all $i > i_{\mathsf{mid}}(T_X)$, the challenge

$$\mathsf{NPC}\left((X'_\alpha)_{v_i(X'_\alpha, Y'_\beta)}, (Y'_\beta)_{q_{\mathsf{obs}}(X'_\alpha, Y'_\beta)}\right) \quad (4)$$

is fixed. To this end we show that

$$\mathsf{NPC}\left((X'_\alpha)_{v_i(X'_\alpha, Y_\beta)}, (Y_\beta)_{q_{\mathsf{obs}}(X'_\alpha, Y_\beta)}\right)$$

is a deterministic function of Y_β. Indeed, in such case and since the challenge consists of a relatively small number of bits, we can find a low-deficiency subsource $Y'_\beta \subset Y_\beta$ such that the random variable in Equation (4) is fixed to a constant. For $i > i_{\mathsf{mid}}(T_X)$, our starting point is the random variable $\mathsf{NPC}\left((X_\alpha)_{v_i(X_\alpha, Y_\beta)}, (Y_\beta)_{q_{\mathsf{obs}}(X_\alpha, Y_\beta)}\right)$. To make this variable depend solely on Y_β, by moving to a subsource of X_α, we must consider the 3 appearances of X_α. We start with $q_{\mathsf{obs}}(X_\alpha, Y_\beta)$.

Claim 1. There exists a deficiency $\log n$ subsource $X'_\alpha \subset X_\alpha$ such that $q_{\mathsf{obs}}(X'_\alpha, Y_\beta)$ is fixed.

PROOF. Let $i_{\mathsf{bot}}(T_Y)$ denote the depth of $u_{\mathsf{bot}}(T_Y)$. Recall that the path $q_{\mathsf{obs}}(X_\alpha, Y_\beta)$ contains the entropy-path of T_Y. In particular, the nodes $u_0(X_\alpha, Y_\beta), \ldots, u_{i_{\mathsf{bot}}(T_Y)}(X_\alpha, Y_\beta)$ are fixed. It is left to argue that there is a low-deficiency subsource $X'_\alpha \subset X_\alpha$ such that all of the remaining nodes, namely, $u_{i_{\mathsf{bot}}(T_Y)+1}(X'_\alpha, Y_\beta), \ldots, u_{\log(n)-1}(X'_\alpha, Y_\beta)$ are fixed as well.

Let us first consider the random node $u_{i_{\mathsf{bot}}(T_Y)+1}(X_\alpha, Y_\beta)$ that is the son of the fixed node $u_{i_{\mathsf{bot}}(T_Y)}(X_\alpha, Y_\beta) = u_{\mathsf{bot}}(T_Y)$. According to Equation (2), the node $u_{\mathsf{bot}}(T_Y)$ decides which of its two sons its favor, namely, which of its sons will be on $q_{\mathsf{obs}}(X_\alpha, Y_\beta)$, according to whether or not

$$\mathsf{Resp}\left((Y_\beta)_{u_{\mathsf{bot}}(T_Y)}, X_\alpha, \mathsf{GSC}\left((Y_\beta)_{\mathsf{leftSon}(u_{\mathsf{bot}}(T_Y))}, X_\alpha\right)\right) = \mathsf{hasEntropy}. \quad (5)$$

By definition, $u_{\mathsf{bot}}(T_Y)$ is a descendant of $\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))$. Further, $(Y_\beta)_{\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))}$ is fixed. Thus, also $(Y_\beta)_{u_{\mathsf{bot}}(T_Y)}$ and $(Y_\beta)_{\mathsf{leftSon}(u_{\mathsf{bot}}(T_Y))}$ are fixed to some constants. Therefore, the Boolean expression in Equation (5) is a deterministic function of X_α. One can show that there exists a deficiency 1 subsource X' of X_α such that the Boolean expression in Equation (5) is fixed. In particular, $u_{i_{\mathsf{bot}}(T_Y)+1}(X', Y_\beta)$ is fixed to a constant.

At this point we can apply the same argument to $i_{\mathsf{bot}}(T_Y)+2$. Indeed, $u_{i_{\mathsf{bot}}(T_Y)+1}(X', Y_\beta)$ is fixed to a constant and all appearances of Y_β in the Boolean expression that is analogous to Equation (5) are again fixed to constants for the same reason as before. Since this process terminates after at most $\log n$ steps and since in each iteration we move to a deficiency 1 subsource of the previous obtained subsource, the claim follows. □

We turn to show that for all $i > i_{\mathsf{mid}}(T_X)$,

$$\mathsf{NPC}\left((X'_\alpha)_{v_i(X'_\alpha, Y_\beta)}, (Y_\beta)_{q_{\mathsf{obs}}(X'_\alpha, Y_\beta)}\right)$$

is a deterministic function of Y_β. By the discussion above, this will prove the first part of Proposition 1. By Claim 1, we already know that $q_{\mathsf{obs}}(X'_\alpha, Y_\beta)$ is fixed to a constant. Thus, it suffices to show that $(X'_\alpha)_{v_i(X'_\alpha, Y_\beta)}$ is a deterministic function of Y_β for all $i > i_{\mathsf{mid}}(T_X)$. By an argument similar to the one used in the proof of Claim 1, one can show that for any such i, $v_i(X'_\alpha, Y_\beta)$ is a deterministic function of Y_β. Note further that, by the definition of an entropy-tree, since $i > i_{\mathsf{mid}}(T_X)$, we have that $v_i(X'_\alpha, Y_\beta)$ is always (that is, for every $(x,y) \in \sup((X'_\alpha, Y_\beta)))$ a descendant of $\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))$. Since $(X'_\alpha)_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))}$ is fixed to a constant we conclude that $(X'_\alpha)_{v_i(X'_\alpha, Y_\beta)}$ is indeed a deterministic function of Y_β.

By the discussion above, we are now in a position to obtain a low-deficiency subsource $Y'_\beta \subset Y_\beta$ such that the random variable $\mathsf{NPC}\left((X'_\alpha)_{v_i(X'_\alpha, Y'_\beta)}, (Y'_\beta)_{q_{\mathsf{obs}}(X'_\alpha, Y'_\beta)}\right)$ is fixed to a constant. We can then apply the challenge-response mechanism and conclude that there exist low-deficiency subsources $X_{\alpha,\beta} \subset X'_\alpha$, $Y_{\alpha,\beta} \subset Y'_\beta$ such that for any $(x,y) \in \sup((X_{\alpha,\beta}, Y_{\alpha,\beta}))$, it holds that

$$\forall i > i_{\mathsf{mid}}(T_X) \quad \mathsf{Resp}\left(x, y, \mathsf{NPC}\left(x_{v_i(x,y)}, y_{q_{\mathsf{obs}}(x,y)}\right)\right) = \mathsf{fixed}.$$

The challenge of $v_{\mathsf{mid}}(T_X)$ is left unresponded

To prove Proposition 1, it suffices to show that w.h.p over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$, it holds that

$$\mathsf{Resp}\left(x, y, \mathsf{NPC}\left(x_{v_{\mathsf{mid}}(T_X)}, y_{q_{\mathsf{obs}}(x,y)}\right)\right) = \mathsf{hasEntropy}.$$

As $u_{\mathsf{top}}(T_Y)$ is on the path $q_{\mathsf{obs}}(x,y) \quad \forall (x,y) \in \sup(X_{\alpha,\beta}, Y_{\alpha,\beta})$, the matrix $\mathsf{NPC}\left((X_{\alpha,\beta})_{v_{\mathsf{mid}}(T_X)}, (Y_{\alpha,\beta})_{q_{\mathsf{obs}}(X_{\alpha,\beta}, Y_{\alpha,\beta})}\right)$ contains the row

$$\mathsf{BExt}\left((Y_{\alpha,\beta})_{u_{\mathsf{top}}(T_Y)}, (X_{\alpha,\beta})_{v_{\mathsf{mid}}(T_X)}\right). \quad (6)$$

Since $X_{v_{\mathsf{mid}}(T_X)}$ is a block-source, $(X_\alpha)_{v_{\mathsf{mid}}(T_X)}$ has a significant amount of entropy. Indeed, X_α is obtained from X by fixing $X_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))} = \mathsf{left}(X_{v_{\mathsf{mid}}(T_X)})$. As $X_{\alpha,\beta}$ is a low-deficiency subsource of X_α, $(X_{\alpha,\beta})_{v_{\mathsf{mid}}(T_X)}$ also has a significant amount of entropy.

We now observe that $(Y_{\alpha,\beta})_{u_{\mathsf{top}}(T_Y)}$ is a block-source. Indeed, $Y_{u_{\mathsf{top}}(T_Y)}$ is a block-source and Y_β is obtained from Y by fixing $Y_{\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))}$. Since $Y_{u_{\mathsf{mid}}(T_Y)}$ is a block-source, this fixing leaves some entropy in $(Y_\beta)_{u_{\mathsf{mid}}(T_Y)}$. Recall further that $(Y_\beta)_{u_{\mathsf{mid}}(T_Y)}$ lies inside $\mathsf{left}((Y_\beta)_{u_{\mathsf{top}}(T_Y)})$ as $u_{\mathsf{mid}}(T_Y)$ is a descendant of $\mathsf{leftSon}(u_{\mathsf{top}}(T_Y))$. Thus, $(Y_{\alpha,\beta})_{u_{\mathsf{top}}(T_Y)}$ is a block-source.

Consider now any low-deficiency subsources $\hat{X} \subset X_{\alpha,\beta}$, $\hat{Y} \subset Y_{\alpha,\beta}$. One can show that $\hat{X}_{v_{\mathsf{mid}}(T_X)}$ has a significant amount of entropy and that $\hat{Y}_{u_{\mathsf{top}}(T_Y)}$ is a block-source (with some deterioration in parameters). Thus, for any low-deficiency subsources $\hat{X}, \hat{Y}$ of $X_{\alpha,\beta}$, $Y_{\alpha,\beta}$, respectively, we have that the challenge matrix associated with $v_{\mathsf{mid}}(T_X)$ contains a row that is close to uniform. In particular this matrix is close to having high entropy. Thus, by the challenge-response mechanism, we have that the node-path challenge associated with $v_{\mathsf{mid}}(T_X)$ is left unresponded with high probability over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$, as desired.

2.5 Determining The Output

Lastly, we compute the output

$$\mathsf{Disp}(x,y) = \mathsf{BExt}\left(x_{v^{\mathsf{obs}}_{\mathsf{mid}}(x,y)} \circ x, y\right),$$

where by $x_{v^{\mathsf{obs}}_{\mathsf{mid}}(x,y)} \circ x$ we denote the block-source with first block $x_{v^{\mathsf{obs}}_{\mathsf{mid}}(x,y)}$ and second block that equals x. There are two potential problems with applying BExt the way we do above. First, we see that the block-source fed to BExt depends on the sample y, which is problematic since y is used as a sample from the source as well. This, however, is a non-issue. Indeed, recall that with high probability over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$ it holds that $v^{\mathsf{obs}}_{\mathsf{mid}}(x,y) = v_{\mathsf{mid}}(T_X)$, and so ignoring a small error, the computation of the extractor BExt above is the same as $\mathsf{BExt}\left(x_{v_{\mathsf{mid}}(T_X)} \circ x, y\right)$.

Now that we have shown that there are no dependencies between the two samples fed to BExt, we only need to make sure that the first sample is indeed coming from a block-source when sampling $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$. Too see why this is true, recall that $v_{\mathsf{mid}}(T_X)$ is a descendant of $\mathsf{leftSon}(v_{\mathsf{top}}(T_X))$ and that $X_{v_{\mathsf{top}}(T_X)}$ is a block-source. As $X_{\alpha,\beta}$ is obtained from X by fixing $X_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))}$ (and by moving to low-deficiency subsources) and since $X_{v_{\mathsf{mid}}(T_X)}$ is a block-source, we have that $(X_{\alpha,\beta})_{v_{\mathsf{top}}(T_X)}$ is also a block-source. Therefore, $(X_{\alpha,\beta})_{v_{\mathsf{mid}}(T_X)} \circ X_{\alpha,\beta}$ is also a block-source. This shows that the application of BExt above is valid, and the output is close to uniform with high probability over $(X_{\alpha,\beta}, Y_{\alpha,\beta})$. In particular, the output is non-constant.

3. ACKNOWLEDGEMENT

This work was done while the author was at the Department of Computer Science and Applied Mathematics, Weiz-

mann Institute of Science. I wish to thank Ran Raz and Avi Wigderson for their warm encouragement.

On a personal note, it is uncustomary to acknowledge one's partner in life in mathematical papers. However, given that this paper was intensively written in the last month of my wife's pregnancy and in the first month of parenthood to the newborn baby girl Meshi and to our sweet Yahli, I will allow myself to make an exception – thank you Orit! Your support and belief in my abilities are uncanny.

4. REFERENCES

[Abb72] H. L. Abbott. Lower bounds for some Ramsey numbers. *Discrete Mathematics*, 2(4):289–293, 1972.

[Alo98] N. Alon. The shannon capacity of a union. *Combinatorica*, 18(3):301–310, 1998.

[Bar06] B. Barak. A simple explicit construction of an $n^{\tilde{o}(\log n)}$-Ramsey graph. *arXiv preprint math/0601651*, 2006.

[BKS+10] B. Barak, G. Kindler, R. Shaltiel, B. Sudakov, and A. Wigderson. Simulating independence: New constructions of condensers, Ramsey graphs, dispersers, and extractors. *Journal of the ACM (JACM)*, 57(4):20, 2010.

[BRSW12] B. Barak, A. Rao, R. Shaltiel, and A. Wigderson. 2-source dispersers for $n^{o(1)}$ entropy, and Ramsey graphs beating the Frankl-Wilson construction. *Annals of Mathematics*, 176(3):1483–1544, 2012.

[CG88] B. Chor and O. Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM Journal on Computing*, 17(2):230–261, 1988.

[Chu81] F.R.K. Chung. A note on constructive methods for Ramsey numbers. *Journal of Graph Theory*, 5(1):109–113, 1981.

[CZ15] E. Chattopadhyay and D. Zuckerman. Explicit two-source extractors and resilient functions. *Electronic Colloquium on Computational Complexity (ECCC)*, 2015.

[Erd47] P. Erdős. Some remarks on the theory of graphs. *Bulletin of the American Mathematical Society*, 53(4):292–294, 1947.

[FN93] A. Fiat and M. Naor. Implicit $O(1)$ probe search. *SIAM Journal on Computing*, 22(1):1–10, 1993.

[Fra77] P. Frankl. A constructive lower bound for some Ramsey numbers. *Ars Combinatoria*, 3:297–302, 1977.

[FW81] P. Frankl and R. M. Wilson. Intersection theorems with geometric consequences. *Combinatorica*, 1(4):357–368, 1981.

[GKRTS05] R. Gradwohl, G. Kindler, O. Reingold, and A. Ta-Shma. On the error parameter of dispersers. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 294–305. Springer, 2005.

[Gop14] P. Gopalan. Constructing Ramsey graphs from Boolean function representations. *Combinatorica*, 34(2):173–206, 2014.

[Gro01] V. Grolmusz. Low rank co-diagonal matrices and Ramsey graphs. *Journal of combinatorics*, 7(1):R15–R15, 2001.

[GS08] A. Gabizon and R. Shaltiel. Increasing the output length of zero-error dispersers. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 430–443. Springer, 2008.

[Li13] X. Li. Extractors for a constant number of independent sources with polylogarithmic min-entropy. In *IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 100–109, 2013.

[Li15a] X. Li. Improved constructions of two-source extractors. *Electronic Colloquium on Computational Complexity (ECCC)*, 2015.

[Li15b] X. Li. Three-source extractors for polylogarithmic min-entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 2015.

[Nag75] Zs. Nagy. A constructive estimation of the Ramsey numbers. *Mat. Lapok*, 23:301–302, 1975.

[Nao92] M. Naor. Constructing Ramsey graphs from small probability spaces. *IBM Research Report RJ 8810*, 1992.

[PR04] P. Pudlák and V. Rödl. Pseudorandom sets and explicit constructions of Ramsey graphs. *Quad. Mat*, 13:327âĂŞ–346, 2004.

[Ram28] F. P. Ramsey. On a problem of formal logic. *Proceedings of the London Mathematical Society*, 30(4):338–384, 1928.

[Yao81] A. Yao. Should tables be sorted? *Journal of the ACM (JACM)*, 28(3):615–628, 1981.