

# Research Statement

Klim Efremenko

October 30, 2011

## 1 Background

My research areas are theoretical computer science and abstract algebra and the interaction between them. Most of my study is focused on understanding the algebraic structure behind combinatorial objects.

The primary focus of my research is Locally Decodable Codes. A code  $\mathcal{C}$  is said to be *Locally Decodable Code* (LDC) with  $q$  queries if it is possible to recover any symbol  $x_i$  of a message  $x$  by making at most  $q$  queries to  $\mathcal{C}(x)$ , such that even if a constant fraction of  $\mathcal{C}(x)$  is corrupted, the decoding algorithm returns the correct answer with high probability.

The main reason that LDCs are important is not because of their obvious applications to data transmission and data storage, but because of their applications to complexity theory and cryptography. Many important results in these fields rely on LDCs. LDCs are closely related to such subjects as worst case – average case reductions, pseudo-random generators, hardness amplification, and private information retrieval schemes, see for example [PS94, Lip90, CKGS98, STV01, Tre03, Tre04, Gas04]. Locally Decodable Codes also found applications in data structures and fault tolerant computations, see for example [CGdW10, dW09, Rom06].

Locally Decodable Codes implicitly appeared in the PCP literature already in the early 1990s, most notably in [BFLS91, PS94, Sud92]. However the first formal definition of LDCs was given by Katz and Trevisan [KT00] in 2000. Since then LDCs became widely used. The first constructions of LDCs [BIK05, KT00] were based on polynomial interpolation techniques. Later on more complicated recursive constructions were discovered [BIKR02, WY07]. All these constructions had exponential length. The tight lower bound of  $2^{\Theta(n)}$  codes were given in [KdW04, GKST06] for two queries LDCs. For many years it was conjectured (see [Gas04, Gol05]) that LDCs should have an exponential dependence on  $n$  for any constant number of queries, until Yekhanin's breakthrough [Yek08]. Yekhanin obtained 3-query LDCs with sub-exponential length. Yekhanin's construction is based on an unproven but a highly believable conjecture in number theory and is quite complicated.

## 2 Current Results

**Locally Decodable Codes:** During my research I attempt to understand the algebraic nature of LDCs.

We say that  $\{u_i\}_{i=1}^k \subset \mathbb{Z}_m^h$  are *S-Matching Vectors* (MV), where  $S$  is a subset of  $\mathbb{Z}_m$  if  $\langle u_i, u_j \rangle \in S \Leftrightarrow i \neq j$ . In [Efr09] based on [Yek08, Rag07] I show that one can construct LDCs from any MV construction in the following theorem:

**Theorem 2.1** ([Efr09]). *For every  $S$ -Matching Vectors  $\{u_i\}_{i=1}^k \subset \mathbb{Z}_m^h$  there exists  $|S| + 1$  query LDC  $\mathcal{C} : \mathbb{F}^k \rightarrow \mathbb{F}^{m^h}$ .*

Applying this result to Grolmusz’s [Gro00] construction gives

**Theorem 2.2** ([Efr09]). *For every  $r$  and for every  $k$  there exists a  $2^r$  query LDC  $\mathcal{C} : \mathbb{F}^k \rightarrow \mathbb{F}^n$ , where  $n = \exp(\exp(O(\sqrt[r]{\log n}(\log \log n)^{r-1})))$ .*

This gives the first unconditional construction of LDCs of sub-exponential length. This construction greatly simplifies Yekhanin’s construction, eliminates the dependence on number theoretic conjectures and improves the parameters of the code.

The history of Matching Vectors is similar to the history of LDCs. It was conjectured for many years that there must be a polynomial upper bound on the size of MV, until Grolmusz’s [Gro00] breakthrough, which showed super-polynomial lower bounds. However, we are still far from the understanding the best possible parameters for LDCs and MVs. There is still not even a conjecture about this. Therefore, although the construction in [Efr09] is pretty simple, it still did not explain the real nature of the LDCs. This led us to seek a new approach to understanding LDCs. In [Efr11] I started a systematic study of LDCs from the point of view of representation theory. In [Efr11] a tight connection between LDCs and irreducible representations is shown in the following theorem:

**Theorem 2.3** ([Efr11]). *Let  $G$  be a finite group and let  $(\rho, \mathbb{F}^k)$  be an irreducible representation of a group  $G$  and  $q$  elements  $g_1, g_2, \dots, g_q$  in  $G$  such that some linear combination of the matrices  $\rho(g_i)$  is a rank one matrix. Then there exists a  $q$  query LDC  $\mathcal{C} : \mathbb{F}^k \rightarrow \mathbb{F}^G$ .*

Although not trivial, one can show [Efr11] that this approach generalizes Theorem 2.1 and gives what we believe is the real algebraic nature behind LDCs. Although the question of how to construct such representations is a very natural one, it was never considered before. I believe that this study of LDCs will bring many natural and interesting questions to the representation theory and that both fields will benefit from this study.

**Pattern Matching:** The most classical problem in *pattern matching* is: given a text of length  $n$  and given a pattern of length  $m$  find all occurrences of the pattern in the text. An almost optimal solution for this problem was already found in the 1970s. This problem has two natural generalizations: the first one is when the text and the pattern contain special symbols called “don’t cares” which match all other symbols; the second one is when we want to find all occurrences of the pattern with at most  $k$  mismatches. For each of these problems individually there exists a good solution, but unfortunately there was no efficient solution which could handle both of these generalizations at the same time. In [CEPR10] we show a randomized algorithm for the  $k$  mismatch problem with “don’t cares” that runs in time  $\tilde{O}(nk)$ . Our approach is based on a sampling technique. We show how to sample one random mismatch from the set of all mismatches for every offset of the pattern in time  $\tilde{O}(n)$ . Running the sampling algorithm  $\tilde{O}(k)$  times gives us a *randomized* algorithm for the  $k$  mismatch problem with “don’t cares”. Later on in [CEPR09] we found a close similarity between the sampling technique and efficient encoding and decoding algorithms for Reed-Solomon codes of distance 3. Using the tools developed for efficient encoding and decoding of Reed-Solomon codes, we gave a *deterministic* algorithm for this problem that whose running time is almost as good as that of the randomized algorithm.

The importance of the sampling technique comes not only from its applications to the problem of  $k$  mismatch, but also from the fact that it could be used in other variants of pattern matching.

For example in [EP08] we use this technique for the following problem: assume that all symbols from the pattern and the text come from some metric space. The question is how fast we can approximate the sum of the distances between symbols of the pattern and the text for all possible offsets of the pattern. In order to solve this problem we use a sampling technique and tools from computational geometry.

**Random Walks:** A random walk on a graph is a process that explores the graph in a random way: at each step the walk is at a vertex of the graph, and at each step it moves to a uniformly selected neighbour of this vertex. Random walks are extremely useful in computer science and in many other fields. The main parameters of a random walk are the time it takes to visit all nodes in the graph (*cover time*) and the time it takes to reach some specific node in the graph (*hitting time*). A very natural problem is to analyze the behavior of  $k$  independent walks in comparison with the behavior of a single walk. Inspired by [AAK<sup>+</sup>07], we initiate [ER09] a systematic study of multiple random walks. We show that the behavior of random walks highly depends on the starting points. We give lower and upper bounds both on the cover time and on the hitting time of multiple random walks over three alternatives for the starting vertices of the random walks: the worst starting vertices (those which maximize the hitting/cover time), the best starting vertices, and starting vertices selected from the stationary distribution. As a rather surprising corollary of our theorems, we obtain a new bound which relates the cover time and the mixing time of a single random walk.

### 3 Future Research

**Locally Decodable Codes:** Although great progress has been made in the construction of short LDCs this question remains wide open. In the future I plan to continue working on this question. In the paper [Efr11] I propose a new approach for constructing shorter LDCs and I would like to continue this line of research. This paper leaves an open question of how to construct irreducible representations which can be transformed to LDCs. I would like to develop algebraic tools which will allow us to answer this question. I believe that the answer to this question will lead to a better understanding of LDCs.

**Self Correctable Codes:** A related notion to LDCs is that of *Self Correctable Codes* (SCC). Self Correctable Codes are codes which allow to correct any symbol of the *codeword* locally. Essentially by definition, any SCC is also an LDC. The converse is not true. No sub-exponential construction of SCCs with constant number of queries is known. Almost all constructions of SCCs are Reed-Muller codes or variants of it. SCCs play an important role in worst case – average case reductions. Therefore, obtaining new constructions of SCCs is an important open problem. I would like to develop the following approach for constructing SCCs:

One can show that if the code  $\mathcal{C}$  is invariant under the permutations of a two-transitive group and there exists a sparse codeword in the dual code, then this code is SCC. The characterization of all two-transitive groups is known. While there exist a very good understanding of codes invariant under permutations of affine group, see [Sud10], there is almost no understanding of codes invariant under other two-transitive groups. I believe that we can use modular representation theory to understand codes invariant under other two-transitive groups and that this will lead to new interesting families of SCCs.

**Matrix Multiplication:** I would like to work on the question of the complexity of matrix multiplication. Determining the complexity of matrix multiplication is long-standing and one of the most important open algorithmic problems. The best known algorithm for matrix multiplication is given by Coppersmith and Winograd [CW90]. Recently, a new group-theoretic approach for matrix multiplication was presented by Cohn and Umans [CU03] and further developed in [CKSU05]. Although these two approaches look very different, they are currently both stuck at the same combinatorial problems. I believe that understanding the algebraic structure behind these combinatorial problems will lead to new insights in the understanding of the complexity of matrix multiplication.

## References

- [AAK<sup>+</sup>07] N. Alon, C. Avin, M. Koucky, G. Kozma, Z. Lotker, and M. R. Tuttle. Many Random Walks Are Faster Than One. *ArXiv e-prints*, 705, May 2007.
- [BFLS91] László Babai, Lance Fortnow, Leonid A. Levin, and Mario Szegedy. Checking computations in polylogarithmic time. In *STOC*, pages 21–31. ACM, 1991.
- [BIK05] Amos Beimel, Yuval Ishai, and Eyal Kushilevitz. General constructions for information-theoretic private information retrieval. *J. Comput. Syst. Sci.*, 71(2):213–247, 2005.
- [BIKR02] Amos Beimel, Yuval Ishai, Eyal Kushilevitz, and Jean-François Raymond. Breaking the  $o(n1/(2k-1))$  barrier for information-theoretic private information retrieval. In *FOCS*, pages 261–270, 2002.
- [CEPR09] Raphaël Clifford, Klim Efremenko, Ely Porat, and Amir Rothschild. From coding theory to efficient pattern matching. In *SODA*, pages 778–784, 2009.
- [CEPR10] Raphaël Clifford, Klim Efremenko, Ely Porat, and Amir Rothschild. Pattern matching with don’t cares and few errors. *J. Comput. Syst. Sci.*, 76(2):115–124, 2010.
- [CGdW10] Victor Chen, Elena Grigorescu, and Ronald de Wolf. Efficient and error-correcting data structures for membership and polynomial evaluation. In *STACS*, pages 203–214, 2010.
- [CKGS98] Benny Chor, Eyal Kushilevitz, Oded Goldreich, and Madhu Sudan. Private information retrieval. *J. ACM*, 45(6):965–981, 1998.
- [CKSU05] Henry Cohn, Robert D. Kleinberg, Balázs Szegedy, and Christopher Umans. Group-theoretic algorithms for matrix multiplication. In *FOCS*, pages 379–388, 2005.
- [CU03] Henry Cohn and Christopher Umans. A group-theoretic approach to fast matrix multiplication. In *FOCS*, pages 438–449, 2003.
- [CW90] Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. *J. Symb. Comput.*, 9(3):251–280, 1990.
- [dW09] Ronald de Wolf. Error-correcting data structures. In *STACS*, pages 313–324, 2009.
- [Efr09] Klim Efremenko. 3-query locally decodable codes of subexponential length. In *STOC*, pages 39–44, 2009.

- [Efr11] Klim Efremenko. From irreducible representations to locally decodable codes, 2011. Unpublished manuscript.
- [EP08] Klim Efremenko and Ely Porat. Approximating general metric distances between a pattern and a text. In *SODA*, pages 419–427, 2008.
- [ER09] Klim Efremenko and Omer Reingold. How well do random walks parallelize? In *APPROX-RANDOM*, pages 476–489, 2009.
- [Gas04] William I. Gasarch. A survey on private information retrieval (column: Computational complexity). *Bulletin of the EATCS*, 82:72–107, 2004.
- [GKST06] Oded Goldreich, Howard J. Karloff, Leonard J. Schulman, and Luca Trevisan. Lower bounds for linear locally decodable codes and private information retrieval. *Computational Complexity*, 15(3):263–296, 2006.
- [Gol05] Oded Goldreich. Short locally testable codes and proofs (survey). *Electronic Colloquium on Computational Complexity (ECCC)*, (014), 2005.
- [Gro00] Vince Grolmusz. Superpolynomial size set-systems with restricted intersections mod 6 and explicit ramsey graphs. *Combinatorica*, 20(1):71–86, 2000.
- [KdW04] Iordanis Kerenidis and Ronald de Wolf. Exponential lower bound for 2-query locally decodable codes via a quantum argument. *J. Comput. Syst. Sci.*, 69(3):395–420, 2004.
- [KT00] Jonathan Katz and Luca Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *STOC*, pages 80–86, 2000.
- [Lip90] Richard J. Lipton. Efficient checking of computations. In *STACS*, pages 207–215, 1990.
- [PS94] Alexander Polishchuk and Daniel A. Spielman. Nearly-linear size holographic proofs. In *STOC*, pages 194–203, 1994.
- [Rag07] Prasad Raghavendra. A note on yekhanin’s locally decodable codes. *Electronic Colloquium on Computational Complexity (ECCC)*, 2007.
- [Rom06] Andrei E. Romashchenko. Reliable computations based on locally decodable codes. In *STACS*, pages 537–548, 2006.
- [STV01] Madhu Sudan, Luca Trevisan, and Salil Vadhan. Pseudorandom generators without the XOR lemma. *Journal of Computer and System Sciences*, 62(2):236–266, 2001.
- [Sud92] Madhu Sudan. *Efficient Checking of Polynomials and Proofs and the Hardness of Approximation Problems*. PhD thesis, University of California at Berkeley, 1992.
- [Sud10] Madhu Sudan. Invariance in property testing. *Electronic Colloquium on Computational Complexity (ECCC)*, pages 51–51, 2010.
- [Tre03] Luca Trevisan. List-decoding using the xor lemma. In *FOCS*, pages 126–135, 2003.
- [Tre04] Luca Trevisan. Some applications of coding theory in computational complexity. Technical Report 043, Electronic Colloquium on Computational Complexity (ECCC), 2004.

- [WY07] David P. Woodruff and Sergey Yekhanin. A geometric approach to information-theoretic private information retrieval. *SIAM J. Comput.*, 37(4):1046–1056, 2007.
- [Yek08] Sergey Yekhanin. Towards 3-query locally decodable codes of subexponential length. *J. ACM*, 55(1), 2008.