# Information Theory (Spring 2015)
# Final Program Report

### Venkatesan Guruswami and David Tse
### (Organizing Co-Chairs)

## SUMMARY

The problems of communication and computation are intrinsically intertwined. To compute, one has to move information around. To communicate reliably, one needs to compute. Information theory, invented by Claude Shannon in 1948 as the mathematical underpinning of communication engineering, has therefore developed natural ties with computer science in general and the theory of computing in particular. Traditionally, though, information theory has been studied in two distinct communities: the communication engineers (who publish in *IEEE Transactions on Information Theory* and attend the *International Symposium on Information Theory, ISIT*) and the theoretical computer scientists (who publish in and attend conferences such as *FOCS/STOC/SODA/CCC*). One of the principal aims of the Information Theory program at the Simons Institute was to bring excellent people from both of these communities together in one place, to facilitate cross-fertilization of ideas, make progress on long-standing open problems, and formulate new problems of mutual interest and far-reaching potential. Information theory is an extremely broad field, and to keep the scope manageable and ensure a critical mass of researchers with common interests, three broad themes were chosen as the focus of the program: theory and applications of error-correcting codes; information theory in combinatorics and computer science; and information theory in machine learning and big data.

We believe that the program was a success even beyond our optimistic expectations. Particularly noteworthy was the formation of new collaborations between groups who would not have had the chance to work together in a sustained fashion but for the environment of the Simons Institute. Furthermore, existing collaborations were strengthened and broadened in scope, and many interesting and timely projects with a focus that cut across the Electrical Engineering/Computer Science boundary were initiated. The results: decade-old long standing open problems solved, a best paper award in STOC 2016, a best paper award in FOCS 2015, a best paper award in NIPS 2015 (one of two awards out of almost 2000 submissions), and a plenary lecture in ISIT 2015, to name just a few concrete accolades.

## STRUCTURE OF THE PROGRAM

A total of 65 senior visitors, 12 Fellows and 20 graduate students attended the program as long-term visitors; all spent at least one month at the Institute, and many (including all the Fellows) stayed for the entire semester-long duration of the program. About half of the participants came from each of the two communities (EE and CS). From the perspective of most visitors, the constellation of people was sufficiently different from the natural peer group they typically meet at conferences that it was possible to get valuable insights into current research in other sub-communities and the techniques used there.

In order to break down language barriers between the communities and facilitate interactions, the program kicked off with a week-long Boot Camp with 3-hour mini-courses on topics that are familiar in one community but less so in the other. The speakers took the requisite time to gently explain the key ideas and concepts, thus sparking many discussions. During the week after the Boot Camp, each of the Research Fellows gave a short talk on his/her research interests to introduce themselves to other Fellows as well as senior visitors.

The program included three Open Lectures, covering the three central themes of the program, which provided the larger UC Berkeley community with an opportunity to learn about the high-level scientific content and wide-ranging scope of the program. The first lecture, by program organizer Alon Orlitsky, reviewed the Good-Turing estimator developed in the quest to crack the Enigma code in WWII, and discussed recent developments that demonstrate its efficacy in estimating the likelihood of unlikely or even unseen events. The second lecture, by Toniann Pitassi, gave a glimpse of the rich theory that has emerged in the last decade generalizing information theory to the setting where two parties engage in an interactive conversation (compared to Shannon’s one-way communication setting). This lecture highlighted the transformative role played by interactive information theory in the subject of communication complexity (which studies the communication cost of distributed tasks), as well as its applications to data privacy and distributed computing. The third lecture, delivered by Rudiger Urbanke, gave a fascinating tour of the rich set of ideas that have gone into coding theory over several decades, all with the common goal of safeguarding data in the presence of noise with the least redundancy and complexity possible. The talk described how ideas from such diverse areas as algebra, number theory, probability, information theory and statistical physics slowly made it from the blackboard into products, while also outlining some of the exciting challenges that still lie ahead.

The semester was punctuated by three week-long workshops on topics that spanned the three major themes of the program:

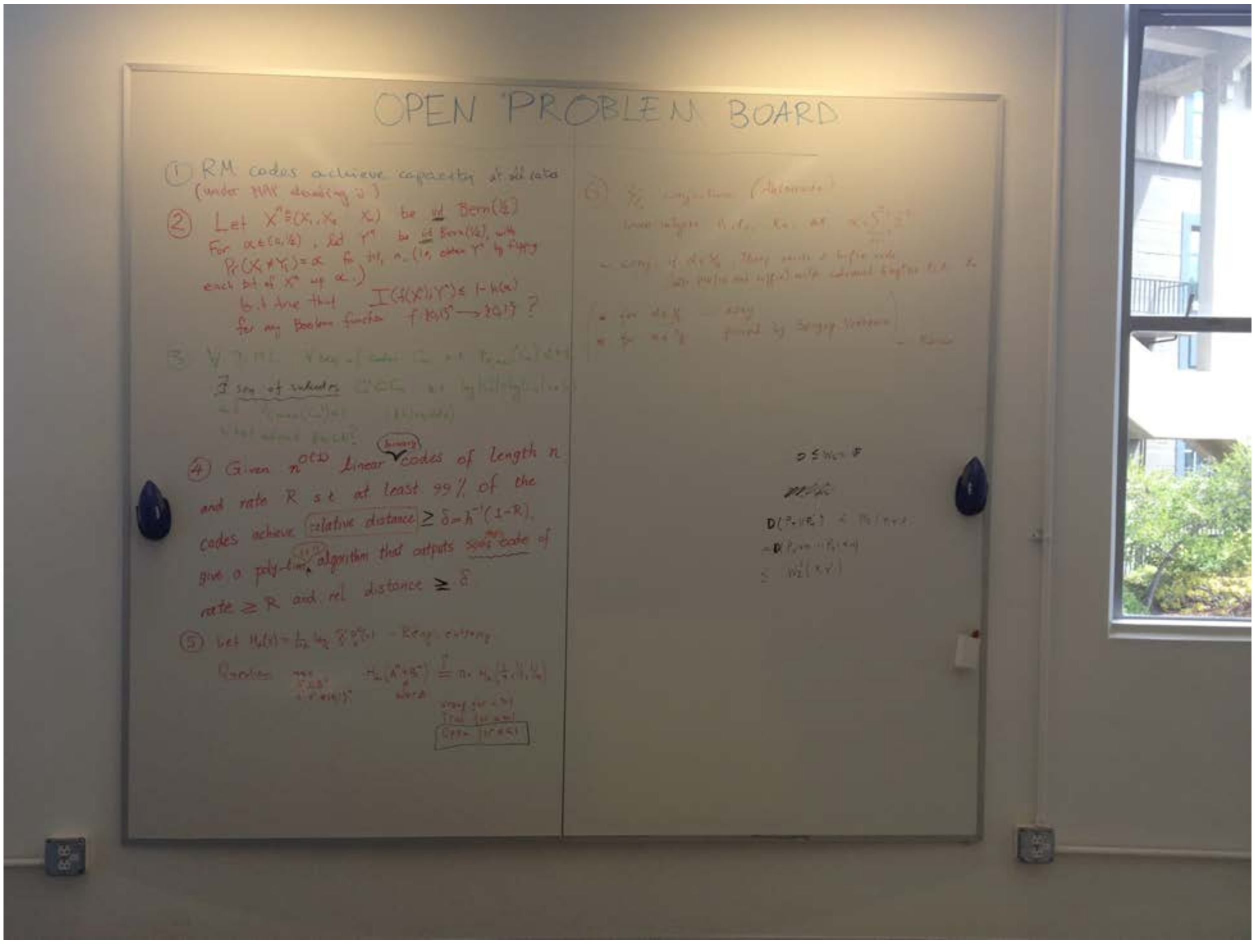
- **Coding Theory:** The explosion in communication and storage demands of the modern era, coupled with many emerging connections of a fundamental nature to computational complexity, big data, signal processing and machine learning, make this an exciting time for coding theory. The aim of this workshop was to highlight coding-theoretic problems emerging in modern applications (such as distributed storage, the upcoming 5G cellular standard, and sequencing applications in biology, among others). In addition to the practical motivations, many of these challenges also have a fundamental flavor. Another goal of the workshop, following the guiding philosophy of the whole program, was to foster stronger interactions and exchange of ideas between the EE-style and CS-style developments in coding theory and applications. Both of these are substantial, with topics such as polar coding and codes for distributed storage drawing the current attention of both communities.

- **Information Theory, Learning and Big Data:** This workshop was centered around the convergence of Information Theory and Machine Learning, driven by the emergence of Big Data as the most important technological and economic engine these days. This nexus of topics is natural and timely, given that recently information theorists have used their toolkit to model and shed light on random phenomena, and machine learning theorists have tried to better understand the fundamental information-theoretic limits of learning from data and exploiting its structure. This workshop brought together a very diverse set of attendees, and had an astonishing 227 registrants.

- **Information Theory in Combinatorics and Complexity Theory:** The focus of this workshop was on the wide-ranging applicability of information-theoretic concepts, intuitions and techniques to the analysis of combinatorial problems arising in computation. Many facets of applications of information theory in complexity theory were represented in the workshop. Central among these was the remarkable progress made recently in understanding information compression in interactive communication protocols, and its application to direct sum theorems and other fundamental results in communication complexity. The workshop featured talks that settled decades-long open questions on deterministic communication complexity (stated, for example, in Chapter 2 of the classical Kushilevitz-Nisan textbook on communication complexity). The role of information theory in parallel repetition (both classical and quantum), extended linear programming formulations and streaming computation were also featured. Applications of information-theoretic ideas to fundamental bounds in combinatorics, especially in additive number theory, boolean function complexity, hypercontractivity, coding theory and geometry were explored in several well-received talks.

Many exciting results from the program, covering all the above themes, were recounted in the reunion workshop in June 2016.

Besides these structured events, the day-to-day routine at the Simons Institute was abuzz with scientific activity. In the non-workshop weeks, participants gathered twice to listen to white-board talks, which sustained the momentum of the program and orchestrated further synergy between the two communities. The whiteboard format was excellent for engaging the audience; there was typically lots of discussion both during the talk and at the coffee break right after it. Spurred by one of these discussions about Reed-Muller codes early in the program, it was decided in an impromptu manner to start an Open Problem Board (see figure below). The first problem listed was whether Reed-Muller codes achieve the channel capacity for all rates. As we will report later, this problem was solved (for the case of erasures) in spectacular fashion during the program.



There were extensive collaborations continually taking place in the numerous excellent open spaces and offices of Calvin Lab. Often these meetings ran for hours on end, with a tremendous focus. On a personal note, the organizers recall many days when they were in research meetings from morning through evening; it was exciting to experience the joys (and happy fatigue) of spending days immersed in research without distractions. Lunch at the many dining choices a short walk from the Institute provided another informal forum to engage in discussions with a wider set of participants, and gain some broader perspective on the hot questions and directions in the various themes as well as the state of the field and its future.

## RESEARCH HIGHLIGHTS

In the remainder of this report we present a small sample of the important directions in which notable progress was made during the Information Theory program.

**Reed-Muller Codes Achieve Capacity:** There is an old adage in information theory: “Almost all codes are good, except the ones we know of.” Some of the oldest codes we know of are the

Reed-Muller (RM) codes, invented in 1954. A long-standing problem is whether they achieve Shannon capacity. This problem was recently revitalized with the breakthrough invention of polar codes by Erdal Arikan, which have strong connections to RM codes. Motivated by one of the tutorials in the Boot Camp on Polar Codes by Arikan himself, and a result by another long term participant Emmanuel Abbe (with Shpilka and Widgerson) showing that RM codes achieve capacity for the erasure channel at rates asymptotically approaching either 0 or 1, the open problem of whether RM codes achieve capacity at *all* rates generated lot of enthusiasm, and was featured prominently as the first problem on the Open Problem Board.

By the end of the program, two groups involving senior visitors and research fellows (Kumar and Pfister, and Kudekar, Mondelli, Sasoglu and Urbanke) had solved the problem for the important special case of the erasure channel [87][88]. Their joint paper won the best paper award at *STOC* 2016, and was also featured in a plenary talk at *ISIT* 2015. The solution of such a classical problem in a matter of months is already a very exciting outcome for the program. What is perhaps even more inspiring is the nature of the solution, which beautifully combines deep ideas from the two communities, very much in the spirit of the program. From theoretical computer science, it draws on the deep result of Friedgut and Kalai that symmetric monotone boolean functions have sharp thresholds. From the communications engineering literature, it applies the elegant *area theorem* for extrinsic information transfer functions. Moreover, the result is applicable to a much broader class of codes beyond RM codes, and thus is expected to have very broad impact. The Simons Institute provides the perfect collaborative environment that makes this kind of breakthrough possible.

**Most Informative Function Conjecture:** The second question on the Open Problems Board was an appealing conjecture published by Thomas Courtade (long term participant) and Gautham Kumar in 2014, called the ``most informative Boolean function conjecture," which asserts that Boolean functions maximizing the mutual information between a binary input vector and the function's value on a noisy version of the input are *dictator functions*. As a tantalizing conjecture bridging information theory and the analysis of Boolean functions, it was only natural that this piqued the interest of both communities attending the Simons program. While the conjecture remains unresolved, there has been much activity around it, and several groups at the Simons Institute were involved in collaborations surrounding this problem. One such group included theoretical computer scientists T.S.Jayram and Amit Chakrabarti and information theorist Chandra Nair, and proved that a full resolution of the conjecture would improve a fundamental inequality, due to Bobkov, on isoperimetry in the Hamming cube. Program co-organizers Venkatesan Guruswami and Jaikumar Radhakrishnan worked on the amount of communication needed to distill perfect shared randomness based on noisy versions of a random string[68]. A corollary of the protocol they developed in this context implied that the Courtade-Kumar conjecture does not hold when the range of the function is large. The communication lower bounds in this work relied on the notion of *hypercontractivity*. For the erasure channel, the necessary hypercontractive inequality was not known despite being a very natural question, and it did not seem to yield to mimicking the calculations for the error channel case. Fortunately, program participant Chandra Nair had developed an equivalent characterization of hypercontractivity parameters using information-theoretic measures [100].

Using this, he was able to prove the requisite hypercontractive inequalities for the erasure channel. This interaction exemplifies the rich variety of intellectual resources and partnerships that the Simons Institute is able to bring to bear on important research problems.

**Gaussian Interference Channel Capacity and Transportation Inequalities:** Shannon's original information theory pertains to communication from a single sender to a single receiver. Subsequently, the theory has been generalized to some *networks*, where multiple senders simultaneously want to communicate with multiple receivers. However, the problem of determining the *network capacity region,* i.e., the maximal rates of information that can be simultaneously transferred between the senders and the receivers, is open for most networks. One of the most canonical networks is the two-user interference channel, where Sender 1 wants to communicate with Receiver 1, and Sender 2 wants to communicate with Receiver 2, each over a noisy channel but causing interference to each other. What is the capacity region of this channel? This problem was formulated in the 1970's and remains open to this date.

In 1985, Costa claimed to have solved this problem in the Gaussian setting for a particular point on the boundary of the capacity region, but it turned out that his proof was flawed. The problem then remained open for a further 30 years. Chandra Nair pointed out this problem to Yury Polyanskiy and Yihong Wu, who were both Fellows in the program, and explained, based on his work [101], that if they could prove a missing technical lemma in Costa's 1985 paper in 1985 then the problem could be solved. During the program, Polyanskiy and Wu were finally able to solve the problem [118]. Their approach was to first apply the transportation-information inequality of Talagrand to reduce the problem to studying continuity of entropy with respect to the so-called Wasserstein distance from optimal transportation. A particularly interesting feature of the proof is that it does not follow the standard strategy of ``single-letterizing" Fano's inequality (a technique responsible for virtually all known tight impossibility bounds in information theory). Their approach is already stimulating follow-up works.

**Coding for Distributed Storage**: One of the biggest impacts coding theory has had in recent years is in distributed storage. Today, massive amounts of data are stored on the cloud spread across several servers. This data needs to be protected by codes so that the data will be safe even if some number of servers fail. This basic problem can be solved by traditional erasure correcting codes, in particular MDS codes that achieve the optimal trade-off between redundancy and number of erasures corrected. However, in these codes, to repair a single node that may have failed or become temporarily unavailable (a frequent occurrence in deployed systems), one needs to download a lot of data from several servers. A simple code that replicates data a few times is very good in this regard; however, it incurs very large redundancy. Two lines of work have emerged in recent years to achieve both good erasure resilience and efficient repair of a failed node: *locally repairable codes* and *regenerating codes*. The former optimize locality (the number of nodes contacted), and the latter the repair bandwidth (the total data downloaded) when repairing a failed node. During the boot camp, detailed tutorials on both these topics were given by Parikshit Gopalan (a theoretical computer scientist) and Alex Dimakis (an information theorist), who did some of the early pioneering work in this area.

Intrigued and directly inspired by one of the open questions mentioned in Dimakis' tutorials, on whether the ubiquitous Reed-Solomon codes can be repaired with non-trivial bandwidth, program organizer Guruswami and long term visitor Mary Wootters started working on this topic. They showed that, contrary to popular belief, it is possible to recover an erased value in Reed-Solomon codes by downloading much less information from other symbols compared to the traditional recovery method using polynomial interpolation [70]. They developed a general framework for bandwidth-efficient repair schemes, and used it to show an improved recovery scheme for the specific code used in the Facebook Analytics Hadoop cluster.

Michael Luby, a program co-organizer, initiated a fundamental new line of research during the program, in which he formulated a new model for distributed storage [94]. The standard model used in distributed storage coding is a static one, focusing on a single time-shot: when a single node fails, the system needs to repair that node using the data from the other nodes. The model Luby considered is a dynamic model with failures occurring over time, and the system need not repair every single node when it fails but needs only to maintain decodability of the overall data at all times. Luby presented his results in the reunion workshop, with the stunning message that, in his model, standard erasure codes are optimal, and specially designed regenerating codes are not needed. While it is too early to assess the impact of this intriguing result, we are sure that it will lead to lots of debate in the field.

**Learning large-alphabet distributions** is a problem that has received a significant amount of attention across statistics, information theory and CS theory in the past decade. One of the world's top experts on this problem, Alon Orlitsky, was a program co-organizer and his presence drove several new research collaborations in this area, including work on the polynomial approximation technique with research fellow Yihong, and on connecting large-alphabet distribution learning with large-scale hypothesis testing with program co-organizer David Tse. (This latter collaboration led to a $1.2M NSF grant.) Moreover, a paper by Alon Orlitsky and his group won the best paper award at *NIPS* 2016 (one of only two awards out of around 2,000 submissions!)

**Biological applications** were yet another thrust of the program. A good example of that work is the collaboration between long-term participant Tom Courtade, program co-organizer David Tse and research fellow Ilan Shomorony on applying information theory to the design of algorithms for DNA assembly from short reads. A central feature of this work is that, although the assembly problem is NP-hard in general, information theory can be used as a lens to focus on those problem instances where the DNA is uniquely reconstructible from the short reads, and these also happen to be instances on which an efficient algorithm exists. This theory has led to a new assembler for third-generation sequencing technologies, and the paper has now been accepted for review by *Nature Methods*.

**List decoding.** One further example of a collaboration across the EE-CS boundary, and between people who had not worked together before, was the work by long term participant Marco Dalai and program co-organizers Guruswami and Radhakrishnan on improved bounds on zero error list decoding capacity (or equivalently, the size of perfect hash families), which improved upon the previous best bounds due to Arikan (1994) in some cases, and may renew interest in these fundamental questions.

There were several other noteworthy outcomes that do not fall strictly within the scope of the program but had connections with it. Research fellows Ben Rossman and Li-Yang Tan proved (with Rocco Servedio) an average-case depth hierarchy theorem for Boolean circuits, which implied hat the polynomial hierarchy is infinite relative to a random oracle, solving a longstanding open problem in computational complexity theory that was posed in 1986. The resulting paper was submitted to *FOCS* 2015 during the program, and received the Best Paper Award at the conference.

In Summer 2015, immediately following the program, long-term participant David Zuckerman and his student Eshan Chattopadhyay gave a breakthrough construction of two-source extractors. This solved a major decades-old open problem in pseudorandomness and combinatorics, for which they won the *STOC* 2016 best paper award. One of the crucial ingredients used in this construction was a non-malleable code, an exciting recent cryptographically inspired concept that gives a way to code data and protect its integrity in settings where traditional error-detection/correction are impossible. Non-malleable coding was one of the topics discussed intensively during the program, for instance in the detailed seminar given by research fellow Mahdi Cheraghchi on his works with Guruswami on capacity and constructions of non-malleable codes.

Finally, program organizer Guruswami collaborated with Zuckerman on a fundamental and surprisingly overlooked problem: fitting a low-degree polynomial to data when there are outliers (i.e., points whose values are far off from the true value). Traditional least squares fit is very sensitive to outliers, and it was not known how to handle even a tiny fraction of outliers efficiently. In this work, which was initiated during the program and later submitted and accepted to *FOCS* 2016, they adapted ideas from the classical Welch-Berlekamp Reed-Solomon decoding algorithm to correct a reasonable fraction of outliers. Subsequently, Eric Price has managed to give an efficient solution to this fundamental problem even when the fraction of outliers approaches ½, which is clearly the maximum allowable.

## CONCLUSIONS

Four years ago, when we conceived the program, our main goal was to bring together the EE and CS communities interested in Information Theory in order that they could learn from each other and potentially obtain major advances in the subject. We believe we have succeeded beyond our most ambitious expectations, as exemplified by the sequence of breakthrough results discussed above, which emerged during the program or as a direct consequence of it. A secondary goal was to build a stronger connection between theory and practice. We believe this goal was also met, notably by advances in the areas of distributed storage and genome sequencing.

## Information Theory, Spring 2015

[1] Y. ABBASI-YADKORI, W. KOOLEN, A. MALEK, and P. L. BARTLETT. Minimax time series prediction. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pp. 2548–2556, 2015.

[2] N. ALON, M. BRAVERMAN, K. EFREMENKO, R. GELLES, and B. HAEUPLER. Reliable Communication over Highly Connected Noisy Networks. In *Proceedings of ACM Symposium on Principles of Distributed Computing (PODC)*, 2016.

[3] A. ANDONI and I. RAZENSHTEYN. Tight Lower Bounds for Data-Dependent Locality-Sensitive Hashing. In *Proceedings of 32nd International Symposium on Computational Geometry (SoCG)*, 2015.

[4] A. ANDONI, I. RAZENSHTEYN, P. INDYK, L. SCHMIDT, and T. LAARHOVEN. Practical and Optimal LSH for Angular Distance. In *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015.

[5] E. ANGELINO, M. J. JOHNSON, and R. P. ADAMS. Patterns of Scalable Bayesian Inference. *arXiv preprint arXiv:1602.05221,* 2016. Under review at Foundations and Trends in Machine Learning.

[6] E. ARIKAN. Varentropy decreases under the polar transform. *IEEE Transactions on Information Theory,* 62, 6, 2016. In preparation.

[7] E. ARIKAN and J. THATHACHAR. A note on some inequalities used in polarization. Preprint. In submission.

[8] P. AWASTHI, M. CHARIKAR, R. KRISHNASWAMY, and A. K. SINOP. The Hardness of Approximation of Euclidean k-means. In *Symposium on Computational Geometry*, 2015.

[9] S. BALAKRISHNAN, M. J. WAINWRIGHT, and B. YU. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *arXiv preprint arXiv:1408.2156,* 2014. To appear in Annals of Statistics.

[10] R. BARBOSA, A. ENE, H. L. NGUYEN, and J. WARD. The Power of Randomization: Distributed Submodular Maximization on Massive Datasets. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[11] P. L. BARTLETT, W. KOOLEN, A. MALEK, E. TAKIMOTO, and M. WARMUTH. Minimax Fixed-Design Linear Regression. In *JMLR Workshop and Conference Proceedings, Proceedings of the Conference on Learning Theory (COLT2015)*, volume 40, pp. 226–239, 2015.

[12] S. BEIGI, S. LIU, C. NAIR, and M. YAZDANPANAH. Some results on the scalar Gaussian interference channel. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[13] S. BEIGI and C. M. NAIR. Equivalent characterization of reverse Brascamp-Lieb inequalities using information measures. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[14] A. BHASKAR, A. JAVANMARD, T. COURTADE, and D. TSE. General probabilistic models of spatial genetic ancestry with applications to stratification correction in genomewide association studies. Preprint. 2015. Submitted to Nature Genetics (also, appeared at Cold Spring Harbor Lab 2015 ProbGen workshop).

[15] C. BHATTACHARYYA, N. GOYAL, R. KANNAN, and J. PANI. NMF under Heavy Noise. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2016.

[16] J. BRAKENSIEK, V. GURUSWAMI, and S. ZBARSKY. Efficient low-redundancy codes for correcting multiple deletions. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1884–1892, 2016.

[17] M. BRAVERMAN, A. GARG, T. MA, H. L. NGUYEN, and D. P. WOODRUFF. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the 48th ACM Symposium on Theory of Computing (STOC)*, 2016.

[18] M. BRAVERMAN, Y. K. KO, A. RUBINSTEIN, and O. WEINSTEIN. ETH Hardness for Densest-k-Subgraph with Perfect Completeness. *arXiv preprint arXiv:1504.08352,* 2015.

[19] N. L. BRAY, H. PIMENTEL, P. MELSTED, and L. PACHTER. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology,* 34, 5, pp. 525–527, 2016.

[20] G. BRESLER and M. KARZAND. Learning tree-structured Ising models in order to make predictions. *arXiv preprint arXiv:1604.06749,* 2016.

[21] G. BRESLER and S. NEGAHBAN. Statistical computational trade-offs for support recovery via greedy algorithms. In preparation.

[22] G. BRESLER, D. SHAH, and L. VOLOCH. Collaborative filtering with low regret. *arXiv preprint arXiv: 1507.05371 [cs.LG],* 2015.

[23] B. BUKH and V. GURUSWAMI. An improved bound on the fraction of correctable deletions. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1893–1901, 2016. Journal version with Johan HASTAD as co-author under review at IEEE Transactions on Information Theory.

[24] A. CHAKRABARTI and S. KALE. Strong Fooling Sets for Multi-Player Communication with Applications to Deterministic Estimation of Stream Statistics. Preprint. FOCS 2016, to appear.

[25] K. CHANDRASEKARAN, M. CHERAGHCHI, V. GANDIKOTA, and E. GRIGORESCU. Local Testing of Lattices. Submitted. 2015.

[26] A. CHATTOPADHYAY, J. EDMONDS, F. ELLEN, and T. PITASSI. Upper and Lower Bounds on the Power of Advice. Preprint. 2016. Accepted SIAM Journal on Computing.

[27] A. CHATTOPADHYAY and S. MUKHOPADHYAY. Tribes Is Hard in the Message Passing Model. In *32nd International Symposium on Theoretical Aspects of Computer Science (STACS 2015)*, volume 30, pp. 224–237, Schloss Dagstuhl--Leibniz-Zentrum fuer Informatik, 2015.

[28] A. CHATTOPADHYAY and A. RUDRA. The range of topological effects on communication. In *ICALP*, pp. 540–551, 2015.

[29] X. CHEN, I. OLIVEIRA, R. A. SERVEDIO, and L.-Y. TAN. Near-optimal small-depth lower bounds for small-distance connectivity. In *Proceedings of the 48th Annual Symposium on the Theory of Computing (STOC)*, 2016.

[30] Y. CHEN, G. KAMATH, C. SUH, and D. TSE. Community recovery in graphs with locality. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[31] Y. CHEN and C. SUH. Spectral MLE: Top-K rank aggregation from pairwise comparisons. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015. Journal of Machine Learning Research, submitted, 2015.

[32] Y. CHEN and C. SUH. Top-K ranking: An information-theoretic perspective. In *Information Theory Workshop-Fall (ITW), 2015 IEEE*, pp. 212–213, IEEE, 2015.

[33] Y. CHEN, C. SUH, and A. GOLDSMITH. Information recovery from pairwise measurements: A Shannon-theoretic approach. In *Proceedings of 2015 IEEE International Symposium on Information Theory (ISIT)*, 2015. To appear in the IEEE Transactions on Information Theory 2016.

[34] M. CHERAGHCHI. Nearly optimal robust secret sharing. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[35] M. CHERAGHCHI, E. GRIGORESCU, B. JUBA, K. WIMMER, and N. XIE. AC0-MOD2 lower bounds for the Boolean Inner Product. In *Proceedings of the 43rd International Colloquium on Automata, Languages and Programming (ICALP)*, 2016.

[36] M. CHERAGHCHI and P. INDYK. Nearly Optimal Deterministic Algorithm for Sparse Walsh-Hadamard Transform. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2016.

[37] M. COSTA and C. NAIR. Gaussian Z-interference channel: around the corner. In *Proceedings of Information Theory and Applications Workshop (ITA)*, 2016.

[38] T. A. COURTADE. Strengthening the entropy power inequality. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016. Submitted to IEEE Transactions on Information Theory.

[39] D. CULLINA, M. DALAI, and Y. POLYANSKIY. Rate-distance tradeoff for codes above graph capacity. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[40] R. CUMMINGS, K. LIGETT, J. RADHAKRISHNAN, A. ROTH, and Z. S. WU. Coordination Complexity: Small Information Coordinating Large Populations. In *Proceedings of the 7th Annual Innovations in Theoretical Computer Science (ITCS)*, pp. 281–290, 2016.

[41] M. DALAI, V. GURUSWAMI, and J. RADHAKRISHNAN. An improved upper bound on the zero-error capacity with list decoding for the 4/3 channel. Manuscript, in preparation.

[42] M. DALAI and Y. POLYANSKIY. Bounds for Codes on Pentagon and Other Cycles. In *53th Annual Allerton Conference on Communication, Control, and Computing*, 2015.

[43] M. DALAI and Y. POLYANSKIY. Bounds on the Reliability of a Typewriter Channel. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[44] A. DEMBO, A. MONTANARI, and S. SEN. Extremal Cuts of Sparse Random Graphs. *arXiv preprint arXiv: 1503.03923v2,* 2015.

[45] Y. DESHPANDE, E. ABBE, and A. MONTANARI. Asymptotic mutual information for the two-group stochastic block model. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016. Submitted to Information and Inference: A Journal of the IMA.

[46] Y. DESHPANDE and A. MONTANARI. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. In *Proceedings of COLT*, 2015.

[47] J. DING, Y. PERES, and G. RANADE. A Tiger by the Tail: When Multiplicative Noise Stymies Control. In

preparation.

[48] C. DWORK, V. FELDMAN, M. HARDT, T. PITASSI, O. REINGOLD, and A. ROTH. Generalization in Adaptive Data Analysis and Holdout Reuse. In *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015.

[49] C. DWORK, A. SMITH, T. STEINKE, J. ULLMAN, and S. VADHAN. Robust Traceability from Trace Amounts. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS)*, 2015.

[50] C. DWORK, W. SU, and L. ZHANG. Private False Discovery Rate Control. Manuscript. 2015.

[51] K. EFREMENKO. Constant-Rate Coding for Multiparty Interactive Communication Is Impossible. In *Proceedings of the 48th ACM Symposium on Theory of Computing (STOC)*, 2016.

[52] A. ENE and H. L. NGUYEN. Random Coordinate Descent Methods for Minimizing Decomposable Submodular Functions. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[53] M. FALAHATGAR, A. JAFARPOUR, A. ORLITSKY, V. PICHAPATI, and A. T. SURESH. Universal Compression of Power-Law Distributions. In *Proceedings of 2015 IEEE International Symposium on Information Theory (ISIT)*, 2015.

[54] F. FARNIA, M. RAZAVIYAYN, S. KANNAN, and D. TSE. Minimum HGR correlation principle: From marginals to joint distribution. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 1377–1381, IEEE, 2015.

[55] F. FARNIA, M. RAZAVIYAYN, and D. TSE. Inference and Feature Selection via Maximal Correlation. In *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015.

[56] V. FELDMAN and J. VONDRAK. Tight Bounds on Low-degree Spectral Concentration of Submodular and XOS functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2015.

[57] U. FERNER, M. MEDARD, and E. SOLJANIN. Why Reading Patterns Matter in Storage Coding & Scheduling Design. *IEEE Cloud,* 2015.

[58] C. FRAGOULI and E. SOLJANIN. (Secure) Linear network coding multicast. *Designs, Codes and Cryptography,* 78, 1, pp. 269–310, 2016.

[59] A. GÁL, E. ALLENDER, and I. MERTZ. Dual VP classes. In *Proceedings of the 40th International Symposium on Mathematical Foundations of Computer Science (MFCS)*, pp. 14–25, 2015.

[60] A. GÁL, M. CHERAGHCHI, and A. MILLS. Correctness and Corruption of Locally Decodable Codes. Manuscript. 2016.

[61] A. GÁL, A. DIMAKIS, A. RAWAT, and Z. SONG. Batch Codes through Dense Graphs without Short Cycles. In *Proceedings of 2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 1477–1481, 2015.

[62] A. GÁL, J. JANG, N. LIMAYE, M. MAHAJAN, and K. SREENIVASAIAH. Space-Efficient Approximations for Subset Sum. Preprint. 2016. To appear in ACM Transactions on Theory of Computing.

[63] A. GÁL and J.-T. JANG. A Generalization of Spira's theorem and circuits with small segragators or separators. Preprint. Submitted for journal publication.

[64] M. GARG and J. RADHAKRISHNAN. Set Membership with a Few Bit Probes. *arXiv preprint arXiv: 1504.02035v1,* 2015.

[65] P. GOPALAN, G. HU, S. SARAF, C. WANG, and S. YEKHANIN. Maximally Recoverable Codes for Grid-like Topologies. *arXiv preprint arXiv:1605.05412v1,* 2016.

[66] P. GOPALAN, R. SERVEDIO, A. TAL, and A. WIGDERSON. Degree and Sensitivity: tails of two distributions. In *Proceedings of the 31st Conference on Computational Complexity (CCC)*, 2016.

[67] V. GURUSWAMI and J. RADHAKRISHNAN. Minimizing the first k shadows of a system of sets. In preparation.

[68] V. GURUSWAMI and J. RADHAKRISHNAN. Tight Bounds for Communication-Assisted Agreement Distillation. In *Proceedings of the 31st Conference on Computational Complexity (CCC)*, pp. 6:1–6:17, 2016.

[69] V. GURUSWAMI and C. WANG. Deletion codes in the high-noise and high-rate regimes. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM*, 2015.

[70] V. GURUSWAMI and M. WOOTTERS. Repairing Reed-Solomon Codes. In *Proceedings of the 48th ACM Symposium on Theory of Computing (STOC)*, 2016.

[71] V. GURUSWAMI and D. ZUCKERMAN. Robust Fourier and Polynomial Curve Fitting. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2016.

[72] J. HAZŁA, T. HOLENSTEIN, and E. MOSSEL. Lower Bounds on Same-Set Inner Product in Correlated Spaces. *arXiv preprint arXiv:1509.06191,* 2015.

[73] J. HAZŁA, T. HOLENSTEIN, and A. RAO. Forbidden Subgraph Bounds for Parallel Repetition and the Density Hales-Jewett Theorem. *arXiv preprint arXiv:1604.05757,* 2016.

[74] P. HARSHA, R. JAIN, and J. RADHAKRISHNAN. Partition bound is quadratically tight for product distributions. In *Proceedings of the 43rd International Colloquium on Automata, Languages, and Programming (ICALP)*, 2016.

[75] N. HARVEY and J. VONDRAK. An algorithmic proof of the Lovasz Local Lemma via resampling oracles. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2015.

[76] S. M. HOSSEIN, T. YAZDI, Y. YUAN, J. MA, H. ZHAO, and O. MILENKOVIC. A Rewritable, Random-Access DNA-Based Storage System. *Scientific Reports,* 5, 14138, 2015.

[77] J. P. HOU, A. EMAD, G. J. PULEO, J. MA, and O. MILENKOVIC. A new correlation clustering method for cancer mutation analysis. *arXiv preprint arXiv:1601.06476,* 2016. To appear in Bioinformatics, 2016.

[78] W. M. HOZA and L. J. SCHULMAN. The adversarial noise threshold for distributed protocols. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 240–258, 2016.

[79] G. JOSHI, E. SOLJANIN, and G. WORNELL. Efficient Redundancy Techniques for Latency Reduction in Cloud Systems. *ACM Trans. on Modeling and Performance Evaluation of Computing Systems,* 2015.

[80] G. JOSHI, E. SOLJANIN, and G. WORNELL. Queues with Redundancy: LatencyCost Analysis. MAMA workshop, SIGMETRICS. 2015.

[81] G. JOSHI, E. SOLJANIN, and G. WORNELL. Using Redundancy to Reduce Latency in Cloud Systems. In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control and Computing*, 2015.

[82] S. KAMATH, A. ORLITSKY, V. PICHAPATI, and A. T. SURESH. On Learning Distributions from their Samples. In *Proceedings of The 28th Conference on Learning Theory*, pp. 1066–1100, 2015.

[83] R. KANNAN. Robust NMF: Going Beyond Separability. Manuscript. 2015.

[84] H. M. KIAH, G. J. PULEO, and O. MILENKOVIC. Codes for DNA Sequence Profiles. *IEEE Trans. Information Theory,* 62, 6, pp. 3135–3146, 2016.

[85] V. KOSTINA. Data compression with low distortion and finite blocklength. In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control and Computing*, 2015.

[86] V. KOSTINA, Y. POLYANSKIY, and S. VERDÚ. Joint Source-Channel Coding with Feedback. In *IEEE Int. Symposium on Information Theory*, 2015.

[87] S. KUDEKAR, S. KUMAR, M. MONDELLI, H. D. PFISTER, E. ŞAŞOĞLU, and R. URBANKE. Comparing the bit-MAP and block-MAP decoding thresholds of Reed-Muller codes on BMS Channels. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[88] S. KUDEKAR, S. KUMAR, M. MONDELLI, H. D. PFISTER, E. ŞAŞOĞLU, and R. URBANKE. Reed-Muller Codes Achieve Capacity on Erasure Channels. In *Proceedings of the 48th ACM Symposium on Theory of Computing (STOC)*, 2016. Subimtted to IEEE IT Transactions.

[89] S. KUDEKAR, T. RICHARDSON, and R. URBANKE. Wave-like solutions of general one-dimensional spatially coupled systems. *arXiv preprint arXiv:1208.5273,* 2012. To appear in IEEE IT Transactions.

[90] S. KUMAR, R. CALDERBANK, and H. D. PFISTER. Beyond Double Transitivity: Capacity-Achieving Cyclic Codes on Erasure Channels. Accepted to ITW. 2016.

[91] S. KUMAR, R. CALDERBANK, and H. D. PFISTER. Reed-Muller Codes Achieve Capacity on the Quantum Erasure Channel. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[92] S. KUMAR and H. D. PFISTER. Reed-Muller Codes Achieve Capacity on Erasure Channels. *arXiv preprint arXiv:1505.05123v2,* 2015.

[93] J. LI, Y. RABANI, L. J. SCHULMAN, and C. SWAMY. Learning Arbitrary Statistical Mixtures of Discrete Distributions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing (STOC)*, pp. 743–752, ACM, 2015.

[94] M. LUBY. A mathematical theory of distributed storage. In preparation.

[95] M. LUBY. Liquid Cloud Storage: large and lazy works best. In preparation.

[96] W. MAO, S. DIGGAVI, and S. KANNAN. Information theory for Nanopore Sequencing. In preparation. Preliminary version at Signal Processing and Communications Conference, Bangalore, 2016.

[97] A. MONTANARI. Finding One Community in a Sparse Graph. *arXiv preprint arXiv:1502.05680v1,* 2015.

[98] A. MONTANARI and S. SEN. Semidefinite Programs on Sparse Random Graphs. *arXiv preprint arXiv: 1504.05910v1,* 2015.

[99] E. MOSSEL and M. I. OHANNESSIAN. On the Impossibility of Learning the Missing Mass. *arXiv preprint arXiv:1503.03613v1,* 2015.

[100] C. M. NAIR and Y. WANG. Evaluation of hypercontractivity parameters using information measures. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[101] C. M. NAIR, L. XIA, and M. YAZDANPANAH. The sub-optimality of the Han-Kobayashi region for the discrete memoryless interference channel. In *Proceedings of 2015 IEEE International Symposium on Information Theory (ISIT)*, 2015.

[102] A. ORLITSKY. Faster Algorithms for Testing under Conditional Sampling. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, 2015.

[103] A. ORLITSKY and A. T. SURESH. Competitive distribution estimation: Why is Good-Turing good. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pp. 2143–2151, 2015.

[104] A. ORLITSKY, A. T. SURESH, and Y. WU. Estimating the number of unseen species: A bird in the hand is worth log n in the bush. *arXiv preprint arXiv:1511.07428v3,* 2016.

[105] A. ÓSKARSDÓTAR, G. MÁSSON, and P. MELSTED. BamHash: a checksum program for verifying the integrity of sequence data. *Bioinformatics,* 2015.

[106] S. OYMAK. Fast and Reliable Estimation from Nonlinear Observations. In preparation.

[107] S. OYMAK. Near-Optimal Bounds for Binary Embedding of Arbitrary Sets. In preparation.

[108] S. OYMAK. Near-Optimal Sample Complexity Bounds for Circulant Binary Embedding. *arXiv preprint arXiv: 1603.03178,* 2016. Short version will appear at ICASSP 2016.

[109] S. OYMAK. On the Universal Phase Transitions of Linear Inverse Problems. In preparation.

[110] S. OYMAK, B. RECHT, and M. SOLTANOLKOTABI. Isometric sketching of any set via the Restricted Isometry Property. Manuscript. 2015.

[111] S. OYMAK, B. RECHT, and M. SOLTANOLKOTABI. Sharp Time-Data Tradeoffs for Linear Inverse Problems. Manuscript. 2015.

[112] A. PANANJADY and T. A. COURTADE. Compressing Sparse Sequences under Local Decodability Constraints. In *Proceedings of 2015 IEEE International Symposium on Information Theory (ISIT)*, 2015. Submitted to IEEE Transactions on Information Theory.

[113] H. D. PFISTER and R. URBANKE. Near-Optimal Finite-Length Scaling for Polar Codes over Large Alphabets. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[114] M. PILANCI and M. J. WAINWRIGHT. Newton Sketch: A Linear-time optimization algorithm with linear-quadratic convergence. *arXiv preprint arXiv:1505.02250 [math.OC],* May 2015. Under review at SIAM Journal on Optimization.

[115] T. PITASSI, B. ROSSMAN, R. SERVEDIO, and L. TAN. Exponential Lower Bounds for Bounded-Depth Frege Proofs. In preparation.

[116] T. PITASSI, B. ROSSMAN, R. A. SERVEDIO, and L.-Y. TAN. Poly-logarithmic Frege depth lower bounds via an expander switching lemma. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pp. 644–657, 2016.

[117] Y. POLYANSKIY. On metric properties of maps between Hamming spaces and related graph homomorphisms. *arXiv preprint arXiv:1503.02779v1,* 2015.

[118] Y. POLYANSKIY and Y. WU. Wasserstein continuity of entropy and outer bounds for interference channels. *arXiv preprint arXiv:1504.04419v1,* 2015. To appear in IEEE Trans on Inf Theory, 2016.

[119] M. RABINOVICH, E. ANGELINO, and M. I. JORDAN. Variational consensus Monte Carlo. In *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015.

[120] J. RADHAKRISHNAN, M. DALAI, and V. GURUSWAMI. A new bound on a family of perfect hash functions. In preparation.

[121] G. RANADE, M. RACZ, V. KOSTINA, and Y. PERES. Rate-limited control capacity of systems with unknown gain. In preparation.

[122] A. RAO and M. SINHA. Simplified Separation of Information and Communication. In preparation.

[123] A. S. RAWAT, Z. SONG, A. G. DIMAKIS, and A. GÁL. Batch Codes Through Dense Graphs Without Short Cycles. *IEEE Trans. Information Theory,* 62, 4, pp. 1592–1604, 2016.

[124] T. RICHARDSON, N. MACRIS, R. EL-KATABH, and R. URBANKE. Analysis of Coupled Scalar Recursions by Displacement Convexity. To be submitted to IEEE IT Transactions.

[125] B. ROSSMAN. Correlation Bounds Against Monotone NC^1. In *Conference on Computational Complexity*, 2015.

[126] B. ROSSMAN. The Average Sensitivity of Bounded-Depth Formulas. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2015.

[127] B. ROSSMAN, R. A. SERVEDIO, and L.-Y. TAN. An average-case depth hierarchy theorem for Boolean circuits. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 1030–1048, 2015.

[128] T. ROUGHGARDEN and O. WEINSTEIN. On the Communication Complexity of Approximate Fixed Points. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 23, pp. 55, 2016.

[129] N. SANTHANAM and V. ANANTHARAM. Agnostic Insurance of Model Classes. *Journal of Machine Learning Research,* Nov 2015.

[130] N. SANTHANAM, V. ANANTHARAM, and W. SZPANKOWSKI. Data derived weak universal compression. In preparation.

[131] L. J. SCHULMAN and A. SINCLAIR. Analysis of a Classical Matrix Preconditioning Algorithm. In *Proceedings of the 47th ACM Symposium on Theory of Computing (STOC 2015)*, 2015.

[132] L. J. SCHULMAN and P. SRIVASTAVA. Stability of Causal Inference. In *Proceedings of the 2016 Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016.

[133] I. SHOMORONY, T. COURTADE, and D. TSE. Do Read Errors Matter for Genome Assembly? In *Proceedings of 2015 IEEE International Symposium on Information Theory (ISIT)*, 2015. Submitted to IEEE Transactions on Molecular, Biological, and Multi-Scale Communications, Shannon Centennial Issue.

[134] I. SHOMORONY, G. M. KAMATH, F. XIA, T. A. COURTADE, and D. N. TSE. Partial DNA Assembly: A Rate-Distortion Perspective. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[135] I. SHOMORONY, S. KIM, T. COURTADE, and D. TSE. Information-Optimal Genome Assembly via Sparse Read-Overlap Graphs. In *Proceedings of the 15th European Conference on Computational Biology (ECCB)*, 2016.

[136] N. SILBERSTEIN and A. GÁL. Optimal combinatorial batch codes based on block designs. *Designs, Codes and Cryptography,* 78, 2, pp. 409–424, 2016.

[137] A. K. SINOP. How to Round Subspaces: A New Spectral Clustering Algorithm. Appeared in ACM-SIAM Symposium on Discrete Algorithms (SODA). 2016.

[138] C. SUH, D. TSE, and J. CHO. To feedback or not to feedback. In *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[139] R. P. TORGHABEH and N. P. SANTHANAM. Community Detection Using Slow Mixing Markov Models. In *IEEE BIG DATA CONFERENCE*, Oct 2015.

[140] A. VELINGKER. Bridging the capacity gap between interactive and one-way communication. In preparation.

[141] S. VERDÚ. Information Theory. Book in preparation.

[142] S. VERDÚ. Non-asymptotic covering lemmas. In *IEEE Information Theory Workshop*, 2015.

[143] S. VERDÚ. α-mutual information. Workshop on Information Theory and Applications, UCSD. February 2015.

[144] Y. WU and P. YANG. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *arXiv preprint arXiv:1504.01227v1,* 2015.

[145] Y. YANG, M. PILANCI, and M. J. WAINWRIGHT. Randomized sketches for kernels: Fast and optimal non-parametric regression. *arXiv preprint arXiv:1501.06195 [stat.ML],* 2015. To appear in Annals of Statistics.

[146] J. ZHANG, A. ORLITSKY, and D. TSE. Optimal Large-scale testing: think globally, act accordingly. Preprint. Submitted to Neural Information Processing Systems, 2016.

[147] Y. ZHANG, M. J. WAINWRIGHT, and M. I. JORDAN. Distributed estimation of generalized matrix rank: Efficient algorithms and lower bounds. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.