LIST OF ABSTRACTS

Wednesday, 22.06.2022

09:50-10:20
Talk: Achieving Coding Gain with Coded Computing
Speaker: Emina Solijanin

We consider distributed computing systems that execute tasks of large jobs in parallel to reduce job service times. Parallel task execution is necessary to satisfy the growing computing demands of machine learning algorithms. Large-scale resource sharing causes random fluctuations in computing times. Thus, some tasks, known as stragglers, take much more time to complete. Task replication and erasure coding have been proposed to curtail the variability in service through diversity. However, with added redundancy, the scale of resource sharing becomes more extensive. The fluctuations in service time (that redundancy is trying to counteract) increase. What is then the gain of adding redundancy? This talk assumes that codes exist for many job types and any code rate and discusses the following questions: Should coding or replication be used? When should redundant tasks be issued, and how many? When should stragglers be canceled/relaunched? How much redundancy should be used for a job?

10:20-10:50
Talk: Rateless Sum-Recovery Codes for Distributed Non-Linear Computations
Speaker: Gauri Joshi

Coded computing has been primarily proposed to handle stragglers in distributed matrix computations and cannot be directly applied to non-linear computing. Many common non-linear computations can be written as a sum of inexpensive non-linear functions (e.g. Taylor series). Based on this observation, we propose a new class of rateless codes called rateless sum-recovery codes whose aim is to recover the sum of source symbols, without necessarily recovering individual symbols. Source symbols correspond to individual inexpensive functions and each encoded symbol is the sum of a subset of source symbols. Encoded symbols are computed in a distributed fashion and for a computation that can be written as a sum of m inexpensive functions, successful sum-recovery is possible with high probability as long as slightly more than m encoded symbols are received.

10:50-11:20
Talk: Straggler-Resilient Coded Federated Learning
Speaker: Alexandre Graell I Amat

We present a straggler-resilient secure aggregation scheme for federated learning. Borrowing ideas from coded distributed computing, the proposed scheme introduces redundancy on the devicesâ€™ data across the network using Shamir's secret sharing. The redundancy is leveraged during the learning phase at the central server to update the global model based on the responses of a subset of the devices, thus providing resiliency to stragglers. The proposed scheme yields information-theoretic privacy on the devices' data against up to a given number of honest-but-curious colluding agents. Further, compared to other schemes in the literature, which

deal with straggling devices and dropouts by ignoring their contribution, it does not suffer from the client-drift problem. For a classification problem on the MNIST dataset and a scenario with 120 devices, the proposed scheme achieves a speed-up factor of 6.6 to 18.7 compared to the state-of-the-art scheme LightSecAgg for an accuracy of 95%.

11:50-12:20

Talk: Coded Privacy-Preserving Computation at Edge Networks
Speaker: Hulya Seferoglu

Multi-party computation (MPC) is promising for designing privacy-preserving machine learning algorithms at edge networks. An emerging approach is coded-MPC (CMPC), which advocates the use of coded computation to improve the performance of MPC in terms of the required number of workers involved in computations. The current approach for designing CMPC algorithms is to merely combine efficient coded computation constructions with MPC. We show that this approach fails short of being efficient, e.g., entangled polynomial codes are not necessarily better than PolyDot codes in MPC setting, while they are always better in coded computation. Motivated by this observation, we propose a new construction; Adaptive Gap Entangled (AGE) codes for MPC setup. We show through analysis and simulations that MPC with AGE codes always perform better than existing CMPC algorithms in terms of the required number of workers as well as computation, storage, and communication overhead.

12:20-12:50

Talk: Private Linear Computation: Algorithms, Bounds, and Applications
Speaker: Alex Sprintson

There are many emerging cloud applications that perform computation remotely for a broad range of purposes. In such applications, it is critical to protect the identity of the data items used in the computation from the cloud operators. In this talk, we focus on the Private Linear Computation (PLC) problem and present several achievability schemes and converse proof techniques. We also discuss applications of PLC in data science and machine learning. Joint work with Anoosheh Heidarzadeh and Nahid Esmati.

14:30-15:00

Talk: Shannon-Style Theorems for Adversarial Channels: When Can One Pack (Exponentially) Many Copies of a Given Pattern?
Speaker: Sidharth Jaggi

The task of reliable communication over noisy channels is both of fundamental import in a variety of applications, and has spawned rich and elegant mathematics over the past 70 years. In his seminal 1948 work, Claude Shannon essentially completely characterized the fundamental limits of communication over channels with random noise.
This work considers “harder” channels – those partially controlled by a malicious jamming adversary – and gives the first precise answer for a “coarser” question; specifically, whether or not any meaningful communication is possible. This is known to be equivalent to the problem of packing shapes in high-dimensional spaces – meaningful communication is possible if and only if exponentially many such shapes can be packed.
We answer the problem in great generality: given general (power) constraints on the transmitter and on the `jammer, and general channel laws, we show that a positive communication rate is possible if and only if a certain efficiently computable convex set (a forbidden region) affiliated with the problem does not completely contain a certain well-studied convex set (the set of so-called completely positive distributions). We demonstrate a sharp dichotomy – either exponentially many shapes can be packed (positive communication rate is possible), or at most a constant number of shapes can be packed (essentially no communication is possible), and precisely characterize which of those two scenarios any given adversarial communication problem falls into. In the process we significantly generalize the classical notions of Gilbert-Varshamov codes, Hamming bound, Plotkin bound, and Elias-Bassalygo bounds to this general setting, demonstrate that these generalize Gilbert-Varshamov codes achieve positive rate if and only if the corresponding generalized Plotkin bound is inactive, and demonstrate adversarial channels for which non-i.i.d. input distributions are necessary to attain positive rate.

Perhaps the main contribution is a lemma of independent interest about long sequences of random variables — the task of generating an arbitrarily long sequence of random variables with most pairs having a joint distribution approximately equaling a pre-specified target is possible if and only if the target is a completely positive distribution. This line of work serves as a stepping stone for a variety of novel results for general adversarial channels, including a precise characterization of the capacity of a very general class of causal adversarial channels.
Based on joint works with Yihan Zhang, Xishi (Nicholas) Wang, Michael Langberg, Anand Sarwate, Amitalok Budkuley, and Andrej Bogdanov.

15:00-15:30
Talk: Approximate Coded Computing
Speaker: Stark Draper

In this talk I describe some of our work on coded matrix multiplication when an approximate result will suffice. We are motivated by potential application in optimization and learning where an exact matric product is not required and one would prefer to get a lower-fidelity result faster. We are also motivated by developing rate-distorion analogs for coded computing and were particularly inspired by a recent JSAIT paper by Jeong et al. "epsilon-coded-computing" wherein the authors show that they can recover an intermediate, approximate, result half-way to exact recovery. In this talk I'll build on that prior work to show how to realize schemes in which there are multiple stages of recovery (more than one) en-route to exact recovery. In analog to successive refinement in rate distortion we term this "successive approximation" coding.

15:30-16:00
Talk: Soft BIBD and Product Gradient Codes
Speaker: Lele Wang

Gradient coding is a coding theoretic framework to provide robustness against slow or unresponsive machines, known as stragglers, in distributed machine learning applications. Recently, Kadhe et al. proposed a gradient code based on a combinatorial design, called balanced incomplete block design (BIBD), which is shown to outperform many existing gradient codes in worst-case adversarial straggling scenarios. However, parameters for which such BIBD constructions exist are very limited. In this paper, we aim to overcome such limitations and construct gradient codes which exist for a wide range of system parameters while retaining the superior performance of BIBD gradient codes. Two such constructions are proposed, one based on a probabilistic construction that relax the stringent BIBD gradient code constraints, and the other based on taking the Kronecker product of existing gradient codes. The proposed gradient codes allow flexible choices of system parameters while retaining comparable error.

Thursday, 23.06.2022

09:30-10:00
Talk: Nonlinear Repair Schemes of Reed-Solomon-Codes
Speaker: Itzhak Tamo

The problem of repairing linear codes, particularly Reed Solomon (RS) codes, has attracted a lot of attention in recent years due to its importance in distributed storage systems. In this problem, a failed code symbol (node) needs to be repaired by downloading as little information as possible from a subset of the remaining nodes. There are examples of RS codes with efficient repair schemes, and some are even optimal. However, these schemes fall short in several aspects; for example, they require a considerable field extension degree, and in particular, they do not work over prime fields. In this work, we explore the power of nonlinear repair schemes of RS codes and show that such schemes are crucial over prime fields, and in some cases, they outperform all linear schemes. Based on joint work with Roni Con.

10:00-10:30
Talk: Low-Latency Storage of Replicated Fragments for Memory Constrained Servers
Speaker: Parimal Parag

We consider the setting of a distributed storage system where a single file is subdivided into smaller fragments of same size which are then replicated with a common replication factor across servers of identical cache size. An incoming file download request is sent to all the servers, and the download is completed whenever the request gathers all the fragments. At each server, we are interested in determining the set of fragments to be stored such that the mean file download time for a request is reduced. We model the fragment download time as an exponential random variable independent and identically distributed for all fragments across all servers and show that the mean file download time can be lower bounded in terms of the expected number of useful servers summed over all distinct fragment downloads. We present deterministic storage schemes that attempt to maximize the number of useful servers.

10:30-11:00
Talk: On/Off Privacy for Online Users and DNA
Speaker: Salim El Rouayheb

Online users are gradually getting more control over their privacy and the data collected about them. What is sometimes overlooked is how correlation can compromise their privacy. For example, a user who wants location privacy during a period of time, has to also worry about hiding his location outside of this period, since teleportation is not yet possible. Also, a user in a social network has to worry about his friends' and family's privacy settings due to them having similar behavior.
To formulate this problem, we focus on an "all-or-nothing" setting for privacy which we refer to as On/Off privacy, in which a user can switch between his/her privacy being On or Off depending on many factors such as internet connection, location, or device. The challenge here is that due to correlation one has to worry about privacy, even when privacy is Off. I will focus on information-theoretic measures for On/Off privacy and describe our initial results in two settings. The first setting deals with correlation and privacy in the context of private information retrieval. The second studies On/Off privacy in designing mechanisms for hiding sensitive genotypes in genomic data.

14:45-15:15
Talk: Classification of Balance Sequences
Speaker: Tuvi Etzion

Balanced sequences and balanced codes have attracted a lot of research in the last seventy years due to their diverse applications in information theory as well as other areas of computer science and engineering. There have been some methods to classify balanced sequences. This work suggests two new different hierarchies to classify these sequences. The first one is based on the largest “l” for which each “l”-tuple is contained the same amount of times in the sequence. This property is a generalization for the property required for de Bruijn sequences. The second hierarchy is based on the number of balanced derivatives of the sequence. Enumeration for each such family of sequences and efficient encoding and decoding algorithms will be discussed.

15:15-15:45
Talk: Sequence Reconstruction Problems for Deletions
Speaker: Han Mao Kiah

The sequence reconstruction problem, introduced by Levenshtein in 2001, considers a communication scenario where the sender transmits a codeword from some codebook and the receiver obtains multiple noisy reads of the codeword. Motivated by modern data storage devices, we focus on the t-deletion channel and report recent related results.
In the first part of the talk, we fix the codebook to be an *(l-1)*-deletion-correcting code of length *n* and study the quantity *N(n,l, t)+1*, the minimum number of distinct channel outputs required for unique reconstruction. Previously, *N(n,l,t)* is known only when *l=1,2* and in this talk, we discuss an asymptotically exact solution for all values of *l* and *t* and present a conjecture on the exact value of *N(n,l, t)* for all values of *n*, *l* and *t*.
In the second part of the talk, we study a variant of the problem where the number of noisy reads *N* is fixed and design reconstruction codes that reconstruct a codeword from *N* distinct noisy reads. We look at two classes of codes: namely, the “constrained shifted VT codes” (CSVT) and “constrained shifted VT codes” (higher order CSVT). For the CSVT, we show that we can uniquely reconstruct a codeword from $n^{t-1}/(t - 1)! + O(n^{t-2})$ distinct reads. The CSVT uses *log(log(n)) +O(1)* redundant bits. For the higher order CSVT and the 2-deletion channel, we show that uniquely reconstruct a codeword from five distinct reads. The higher-order CSVT uses *2*log(n) + O(log(log(n)))* redundant bits.

16:15-16:45
Talk: Biocompatible and Biosafe DNA Data Storage
Speaker: Stephan Lemaire

DNA data storage is an emerging technology that has the potential to replace bulky, fragile and energy-intensive digital data storage media. We have developed a storage strategy, called DNA Drive, that organizes data on long double stranded replicative circular DNA molecules and has unlimited capacity. A random exploratory method, named Random Iterative in-Silico Evolution (RISE), ensures the biosafety of the process by limiting the coding potential of the DNA. Using this approach, we stored two historical texts from the French revolution, the Declaration of the rights of man and of the citizen of 1789 and the Declaration of the rights of woman and of the female citizen published in 1791. The biocompatibility of the DNA Drive enables biological manipulation of the data including low cost copy and edition.
Perhaps the main contribution is a lemma of independent interest about long sequences of random variables — the task of generating an arbitrarily long sequence of random variables with most pairs having a joint distribution approximately equaling a pre-specified target is possible if and only if the target is a completely positive distribution. This line of work serves as a stepping stone for a variety of novel results for general adversarial channels, including a precise characterization of the capacity of a very general class of causal adversarial channels.
Based on joint works with Yihan Zhang, Xishi (Nicholas) Wang, Michael Langberg, Anand Sarwate, Amitalok Budkuley, and Andrej Bogdanov.

16:45-17:15

Talk: Image Coding for Long-term Archiving in Synthetic DNA Polymers

Speaker: Marc Antonini

The efficient storage of digital data is becoming very challenging over the years due to the exponential increase in the generation of data which can't compete with the existing storage resources. Furthermore, the infrequently accessed data can be safely stored for no longer than 10-20 years due to the short life-span of conventional storage devices. To this end, recent studies have proven DNA to be a very promising candidate for the long-term storage of digital data. Several pioneering works have proposed different encoding methods for the specific encoding of images into a quaternary DNA representation while first compressing the image using the classical JPEG standard to reduce the high cost of DNA synthesis. However this type of compression is not optimized with respect to the quaternary DNA code and results in an open-loop workflow.

Inspired by the main workflow of the classical JPEG standard, in this presentation we will focus on a modified version of the JPEG algorithm for the encoding of an image into a quaternary representation based on the DNA alphabet composed by the four nucleotides {A,C,T,G}. The proposed algorithm must respect some biochemical constraints imposed by the synthesis chemistry and the sequencing process for decoding.

This approach showed an improvement in the compression ratio (expressed in input bits per nucleotide) compared to state-of-the-art solutions, without affecting the quality of the decoded image.

Friday, 24.06.2022

09:30-10:00

Talk: Group Testing For Covid-19: What are C_t Values and How to Effectively Use Them?
Speaker: Olgica Milenkovic

We discuss the problem of group testing for Covid-19 testing from a coding-theoretic point of view that accounts for the information provided by the measured C_t values, simplicity of testing and dilution effects. In particular, we will discuss novel testing methods based on semiquantitative tests and extension of Hwang's model for dilution effects. Most results will be illustrated on real-world test data. No prior knowledge in biology is needed to follow the talk.

10:00-10:30

Talk: Coding for the Nanopore DNA Sequencing Channel
Speaker: Emanuele Viterbo

The molecular structure of DNA provides a very stable scaffolding of long strands of nucleotide macromolecules: adenine (A), thymine (T), cytosine (C), guanine (G), which support information transfer across all living organisms. DNA has notable advantages over traditional storage media, including a much higher data density, much longer-term durability, and the availability of copies at no additional cost.
The Oxford Nanopore Technology (ONT) sequencing technology was developed in the last 15 years and became commercially available in 2015. With sequencing error rates of the order of 5-10%, it is currently less reliable than other sequencing technologies, but the relatively low cost, high throughput, and real-time output make it the ideal candidate for storage applications. The single-strand DNA (ssDNA) molecules flow through an array of nanopores mounted on a membrane. An ionic current flowing through each nanopore is measured to produce an analog signal, which is a function of multiple nucleotides inside the nanopore.
The sequencing process, reconstructing the nucleotide sequence from the analog signal, is known as ‘base-calling’. Owing to the random and highly variable speed at which a ssDNA flows through the nanopores, the analog signal suffers from random time-warping distortion, which causes insertion and deletion errors in the base-called sequences. Additional measurement noise is present, which causes substitution errors. Base-calling has improved in recent years by using neural networks trained on natural DNA sequences. However, the application of machine learning for sequencing synthetic DNA will require an ad-hoc approach because of the greater randomness of the stored nucleotide sequences.
Standard error control coding techniques are not well suited for synthetic DNA storage systems owing to the very different nature of the impairments in the ONT read and write processes. The aim of this talk is to present a model for the ONT channel and possible research directions in coding, signal processing, and machine learning techniques that enable reliable DNA storage systems.

10:30-11:00

Talk: Computational and statistical aspects of synthetic biology
Speaker: Zohar Yakhini

I will describe, in general lines, several important synthetic biology research areas that my group is involved in. I will focus on data storage in DNA, on CRISPR and on the use of synthetic oligonucleotide libraries (OLs). For each topic I will describe computational challenges as well as some solutions. In particular I will address efficient composite letter

coding for storage, off-target activity calling and quantification for CRISPR and barcode design for OLs. The talk does not assume knowledge in biology.

11:30-12:00
Talk: DNA Archival Storage, a Bottom Up Approach
Speaker: Thomas Heinis

DNA is an attractive medium for storage of digital data in the cloud, owing to a form factor several orders of magnitude smaller than any other. Key to storing binary data in synthetic in DNA is the translation between the binary representation of digital data to the quaternary domain of DNA. This translation must adhere to constraints imposed by the synthesis and sequencing processes used to write and read respectively. A technological advance in either process changes the constraints and renders current encoding schemes obsolete. In this presentation we present a recipe for taking constraints and producing an appropriate encoding scheme. Such a mechanism allows moving the encoding in lockstep with the technological advances in the underlying processes. We further show a method to understand trade-offs in constraints for a given overhead of bits needed to meet such constraints.

12:00-12:30
Talk: DNA Data Storage Using Photolithographic Synthesis and Error-Correction
Speaker: Reinhard Heckel

Due to its longevity and enormous information density, DNA is an attractive medium for archival storage. The current bottleneck of DNA data storage systems is synthesis. The key idea for breaking this bottleneck pursued in this work is to move beyond the low-error and expensive synthesis employed almost exclusively in today's systems, towards cheaper, potentially faster, but high-error synthesis technologies. We discuss a DNA storage system that relies on light-directed synthesis, which is considerably cheaper than conventional solid-phase synthesis. However, this technology has a high sequence error rate. We demonstrate that even in this high-error regime, reliable storage of information is possible, by developing a pipeline of algorithms for encoding and reconstruction of the information.
Joint work with: R. Grass, P. L. Antkowiak, J. Lietard, M. Z. Darestani, M. M. Somoza, W. J. Stark

14:00-14:30
Talk: Coding for Polymer-based Data Storage
Speaker: Ryan Gabrys

Motivated by polymer-based data-storage platforms that use chains of binary synthetic polymers as the recording media and read the content via tandem mass spectrometers, we propose codes that allow for both unique string reconstruction and correction of multiple mass errors. For this setup, we discuss two approaches. The first relies on formulating the string reconstruction problem in terms of a bivariate polynomial, and results in efficiently encodable and decodable coding schemes. The second approach, which requires far less redundant bits, combines a Catalan string constraint with a series of parity symbols. Afterwards, we consider the related problem of string reconstruction from the union of compositions of mixtures of string prefixes and suffixes.

14:30-15:00
Talk: Interleaving of Generalized Gabidulin Codes
Speaker: Vladimir Sidorenko

We define and analyze a new wide class of interleaved generalized Gabidulin codes over a finite field. Component codes can have different lengths, dimensions, supports, and automorphisms. We show which codes in this class reach the Singleton-type bound for the rank metric. For the proposed codes we give an efficient unique syndrome decoding algorithm based on the skew transform which corrects errors beyond half the code distance.