(Photo Credit: Slava Gerovitch)

Computer Science Conference

October 2018

2018 PRIMES Computer Science Conference

Saturday, October 13

Room 4-370, MIT

2:00 pm: Welcoming Remarks

- Dr. Slava Gerovitch, PRIMES Program Director
- Prof. Srini Devadas, PRIMES Computer Science Section Coordinator

2:10 pm: Session 1

- Sanath Govindarajan and Walden Yan, "Secure image classification with lattice-based fully homomorphic encryption" (mentor William Moses)
- John Kuszmaul, "Verkle trees: Ver(y short Mer)kle trees" (mentor Alin Tomescu)
- Yiming Zheng, "Scaling transaction verifications in cryptocurrencies" (mentor Alin Tomescu)
- Robert Chen, "Aleator: Random beacon via scalable threshold signatures" (mentor Alin Tomescu)

3:15–3:30 pm: break

3:30 pm: Session 2

- David Lu, "XRD: A Scalable messaging system with cryptographic privacy" (mentor Albert Kwon)
- Ethan Mendes and Patrick Zhang, "Maintaining the anonymity of direct anonymous attestations with subverted platforms" (mentor Kyle Hogan)
- Shashvat Srivastava, "AnonStake: An Anonymous proof-of-stake cryptocurrency via zero-knowledge proofs and Algorand" (mentor Kyle Hogan)

4:20-4:30 pm: break

4:30 pm: Session 3

- Michael Gerovitch, Neil Malur, and Hari Narayanan, “The Second Opinion Project: : Leveraging external knowledge databases for additional patient medical options” (mentor Dr. Gil Alterovitz)
- Yingtong Zhao, “Server and interface for patient risk assessment” (mentor Dr. Gil Alterovitz)
- Leo Dong, “Novel feature learning method of gene expression data based on an optimized denoising autoencoder” (mentor Dr. Gil Alterovitz)
- Andrew Zhang, “Antimicrobial resistance prediction using deep convolutional neural networks on whole genome sequence data” (mentor Dr. Gil Alterovitz)

5:35-5:45 pm: break

5:45 pm: Session 4

- Anusha Murali, “A Semi-Supervised dimensionality reduction method to reduce batch effects in genomic data” (mentor Dr. Mahmoud Ghandi, Broad Institute)
- Sanjit Bhat, “Towards efficient methods for training robust deep neural networks” (mentor Dimitris Tsipras)
- Aditya Saligrama and Andrew Shen, “A practical analysis of Rust’s concurrency story” (mentor Jon Gjengset)

6:35 pm: Conference ends

2018 PRIMES CS CONFERENCE ABSTRACTS

SATURDAY, OCTOBER 13
ROOM 4-370

SESSION 1

Sanath Govindarajan and Walden Yan

Secure Image Classification with Lattice-Based Fully Homomorphic Encryption

Mentor: William Moses

Machine learning has become an extremely useful tool for finding patterns in large volumes of data. However, since the heavy computations generally require sending data to an untrusted party, several tasks that deal with private information — such as health records and sensitive emails — cannot make use of ML. Here we present our solution: a scheme using encrypted data on low-bit precision neural networks.

John Kuszmaul

Verkle Trees: Ver(y short Mer)kle trees

Mentor: Alin Tomescu

We present *Verkle Trees*, a bandwidth-efficient version of Merkle Hash Trees (MHTs) built using Vector Commitments. MHTs allow for membership proofs and verification of leaves in the tree with regard to a digest, the root of the tree. Merkle trees are efficient; they can be constructed in linear time and proofs are logarithmically sized. Still, Merkle proofs are large enough that they can entail steep bandwidth costs when sent across the network. Vector Commitment (VC) schemes allow for membership proofs and verification of elements in a vector with regard to a digest, the commitment. Vector Commitment schemes are computationally expensive, taking $O(n^2)$ time to build a VC with n elements, but membership proofs are constant-sized. At a high level, both Vector Commitment schemes and Merkle Trees solve the same problem; they allow the computation and verification of membership proofs with respect to a digest. We combine techniques used in MHTs and VCs to create Verkle Trees, which offer a trade off between computational efficiency and bandwidth (in the form of proof-size). A Verkle Tree with n leaves and branching factor q can be constructed in $O(qn)$ time and proofs are of size $O(\log_q n)$.

Yiming Zheng

Scaling Transaction Verifications in Cryptocurrencies

Mentor: Alin Tomescu

We present a new data structure that can help scale the verification of transactions in cryptocurrencies. In cryptocurrencies, miners check the validity of broadcasted transactions and include them in new blocks. Unfortunately, miners currently have to store large amounts of data to verify these transactions. To address this problem, we propose a new data structure for storing transactions in a cryptocurrency. Our data structure is a binary tree built using a novel cryptographic primitive. In our system, clients must store proofs of their own balances which they broadcast as part of a transaction when they wish to send money. Importantly, unlike previous work, miners in our system only store the 32-byte digest and can validate transactions efficiently against it. Furthermore, miners efficiently update the digest to reflect changes in account balances as new blocks with transactions are created by other miners.

Robert Chen

Aleator: Random beacon via scalable Threshold Signatures

Mentor: Alin Tomescu

We present Aleator, a random beacon built using a scalable threshold signature scheme.

Current random beacons that use threshold signatures are bias-resistant but do not scale past thousands of participants.

To address this, we design and implement a BLS threshold signature scheme that scales to hundreds of thousands of participants, resulting in increased security for our random beacon.

Our main contribution is to speed up Lagrange interpolation.

Previous Lagrange-based threshold signature schemes require $O(k^2)$ time to aggregate a threshold signature, where k is the number of participants.

Our technique speeds this up to $O(k \log^2 k)$ time using Fast Fourier Transform (FFT) and multipoint evaluation techniques.

Importantly, the asymptotic improvement in aggregation time is significant in practice: we can aggregate a threshold signature from 100,000 participants in only 20 seconds.

This is a significant improvement over previous (naive) Lagrange-based techniques, which would take 13 minutes to aggregate the same signature.

SESSION 2

David Lu

XRD: A Scalable Messaging System with Cryptographic Privacy

Mentor: Albert Kwon

Even as end-to-end encrypted communication becomes more popular, private messaging remains a challenging problem due to metadata leakages, such as who is communicating with whom. Most existing systems that hide communication metadata either (1) do not scale easily, (2) incur significant overheads, or (3) provide weaker guarantees than cryptographic privacy, such as differential privacy or heuristic privacy. We present XRD (short for Crossroads), a metadata private messaging system that provides cryptographic privacy, while scaling easily to support more users by adding more servers. At a high level, XRD uses multiple mix networks in parallel with several techniques, including hybrid verifiable shuffling. As a result, XRD can support 2 million users with sub-minute latency using 200 servers.

Ethan Mendes and Patrick Zhang

Maintaining the Anonymity of Direct Anonymous Attestations with Subverted Platforms

Mentor: Kyle Hogan

More than ever, users today wish to keep their identity and other personal information hidden from identity thieves and hackers. This idea of remaining anonymous is especially applicable to the realm of proving or attesting that ones device isn't malicious or that it is part of a specific group. A well-known example is Digital Rights Management. Existing processes utilize the Trusted Platform Module (TPM) to perform these attestations. However, all existing anonymous attestation protocols allow the operating system (a difficult body of code to verify due to its large size) to control the process, which can lead to leaked credentials. We have built on preceding work to present an attestation protocol which reduces the reliance on the host operating system by availing both a secure execution environment provided by Intel Software Guard Extensions (SGX) and a secure means of sending the attestation through the operating system using The Onion Router (TOR). This protocol has a more realistic threat model than any past work, from which a strong notion of anonymity can be achieved with the same functionality provided by all previous methods.

Shashvat Srivastava

AnonStake: An Anonymous Proof-Of-Stake Cryptocurrency via Zero-Knowledge Proofs and Algorand

Mentor: Kyle Hogan

We present AnonStake, the first anonymous Proof-of-Stake cryptocurrency. AnonStake builds on Algorand, a Proof-of-Stake cryptocurrency that features fast block times and

consensus. The anonymous transactions are based on ZeroCash, an anonymous cryptocurrency that uses zero-knowledge proofs to hide transaction details. In Algorand consensus, users are selected to form committees at rates proportional to their wealth. Traditionally, users need to know each others account balances to verify that committees were formed properly. We construct a zero-knowledge proof that allows users to participate in Algorand consensus without revealing their identity or their wealth, thus preserving privacy. The construction uses new cryptographic primitives, such as MiMC, that are designed for applications in zero-knowledge proofs.

SESSION 3

Michael Gerovitch, Neil Malur, and Hari Narayanan

The Second Opinion Project: Leveraging External Knowledge Databases for Additional Patient Medical Options

Mentor: Dr. Gil Alterovitz

In medicine, patients typically receive treatment options from physicians. Yet many increasingly sophisticated knowledge bases are now becoming available online, and patients gain access to other sources of information. In order for the patients to learn about new medical treatment options to complement those presented to them, it is important to enable the flow of information between patients and medical researchers.

Our project looks at a wide range of treatments and examines eligibility criteria for clinical trials represented through identifiers from multiple terminologies. First, we analyze the existing gaps between medical domains by mapping concepts from several terminologies. Next, we use a web database interface for accessing lab and gene information on alternative treatments and interventions. We then apply various methods of machine learning and natural language processing to symptom and intervention data. Lastly, we discuss the ultimate goal of our research: a platform that fosters the flow of information to patients.

Yingtong Zhao

Server and Interface for Patient Risk Assessment

Mentor: Dr. Gil Alterovitz

Early diagnosis is hugely important in the treatment of many diseases, and for some like breast and ovary cancer that may be hereditary, predicting the presence of high-risk genes can help determine a patients need for further exams and treatment. However, current solutions to automate the prediction of this likelihood are usually too technically demanding for patients or small organizations to operate.

We present a two-part solution to this accessibility problem. Our API server bridges existing software with the widespread FHIR medical information standard, which used alone allows reuse of a patient's existing medical data in assessing their genetic risk. Also,

our web interface communicates with the FHIR server to allow end users to directly enter their own data in a web browser to receive an automated, user-friendly evaluation. Both parts additionally feature a simplified deployment process using Docker.

Leo Dong

Novel Feature Learning Method of Gene Expression Data Based on an Optimized Denoising Autoencoder

Mentor: Dr. Gil Alterovitz

Advances in bioinformatics sequencing technology enable low-cost measurements of genome-wide gene expressions in organism tissues. Such gene expression data allows for simultaneous observations of thousands of genes, which is especially useful for cancer research in order to understand gene regulations underlying cancer genealogy and even use it for cancer diagnosis on a molecular level. However, gene expression data also poses challenges for quantitative analyses. Most notably, gene expression data has high dimensions, usually thousands of genes, and small sample sizes, a situation often termed the curse of dimensionality. This research proposes a novel feature learning method for gene expression data in order to reduce the dimensionality of the data. The current state-of-the-art method, ADAGE, is based on denoising autoencoders, a special neural network structure. Our work builds off of the current state-of-the-art method by optimizing the denoising autoencoder structure, resulting in a novel feature learning method called net-DAE. We take the breast cancer as case study and use the exact same dataset (namely METABRIC and TCGA-BRCA) as the original ADAGE paper. We show that our method is able to learn more useful features both computationally and biologically in most cases compared to the current method, boosting the accuracies of cancer subtype classification and other supervised tasks.

Andrew Zhang

Antimicrobial Resistance Prediction Using Deep Convolutional Neural Networks on Whole Genome Sequence Data

Mentor: Dr. Gil Alterovitz

The advent of antibiotics has greatly reduced the risk of dying from untreatable bacterial infections. However, antimicrobial resistance (AMR) threatens the effectiveness of antibiotics against bacteria, and has caused thousands of deaths by previously preventable illnesses. One particular issue regarding AMR is that clinicians have difficulty knowing what antibiotics to prescribe to a patient because they dont know what antibiotics the strain in question is resistant to. Currently, there are two ways for doctors to respond: prescribe a broad spectrum of antibiotics in the hope that one will be effective, or send a bacterial culture to a lab for testing. Both of these solutions are problematic. Prescribing a broad spectrum helps contribute to AMR development, thus only making the problem worse in the long run, and traditional culture testing takes at least two days to complete, in which time the patients condition could significantly worsen. Thus, there is a need

for faster identification of a bacterias resistance. Machine learning, coupled with whole genome sequencing, can provide an answer to this dilemma. Given the improvements in sequencing technology, which have made the process much cheaper and faster, and the vast body of bacterial genomic data that has resulted, machine learning algorithms could be used to quickly predict a bacterias resistances in the field. Previous studies have used genomic information from bacteria with basic machine learning techniques, like k-mers and logistic regression, and have already had encouraging results. We build on their work, and propose a method to determine if a bacteria is resistant to an antibiotic based on its genetic information using deep convolutional neural networks. The deep CNN model achieves an average accuracy of 95% as verified on data from Acinetobacter Baumannii resistance to Carbapenem and Klebsiella Pneumoniae resistance to Ampicillin. The model takes less than 25 minutes to train on a PC with a GPU. Once trained, it takes less than a second to make an AMR prediction. Our model, if used together with the real time genome sequencing machine now already available, could make fast and accurate AMR predictions, gaining precious time to save patients life, and preventing spreading of the resistant bacteria.

SESSION 4

Anusha Murali

A Semi-Supervised Dimensionality Reduction Method to Reduce Batch Effects in Genomic Data

Mentor: Dr. Mahmoud Ghandi, Broad Institute

Gene expression datasets generated using different experimental methods suffer from batch effects. In this study, we present a novel dimensionality reduction method and apply it to reduce batch effects in the Cancer Cell Line Encyclopedia mRNA expression datasets. We formulate our investigation as a constraint optimization problem to find a projection that minimizes the normalized distances between paired samples, and solve it using Lagrange multiplier technique for a special case. We demonstrate a geometric representation of the method with a few examples. Finally, we provide a solution to the general case using an eigendecomposition technique.

Sanjit Bhat

Towards Efficient Methods for Training Robust Deep Neural Networks

Mentor: Dimitris Tsipras

In recent years, several works have shown that neural networks are vulnerable to adversarial examples, i.e., specially crafted inputs that look indistinguishable to humans yet cause misclassification. Adversarial training creates robust models by augmenting the dataset with adversarially perturbed inputs. However, since these inputs require significant computation time, adversarial training can be impractical in real-world applications. In this work, we take a closer look at adversarial training and also explore asynchronous

parallelization approaches. Taken together, these two directions enable comparable robustness on the MNIST dataset to prior art while reducing the training time from several hours to just 8 minutes. Overall, our work moves a step closer to the efficient training of robust deep neural networks.

Aditya Saligrama and Andrew Shen

A Practical Analysis of Rust's Concurrency Story

Mentor: Jon Gjengset

Correct concurrent programs are difficult to write; when multiple threads mutate shared data, they may lose writes, corrupt data, or produce erratic program behavior. While many of the data-race issues with concurrency can be avoided by the placing of locks throughout the code, these often serialize program execution, and can significantly slow down performance-critical applications. Programmers also make mistakes, and often forget locks in less-executed code paths, which leads to programs that misbehave only in rare situations.

Rust is a recent programming language from Mozilla that attempts to solve these intertwined issues by detecting data-races at compile time. Rust's type system encodes a data-structures ability to be shared between threads in the type system, which in turn allows the compiler to reject programs where threads directly mutate shared state without locks or other protection mechanisms. In this work, we examine how this aspect of Rust's type system impacts the development and refinement of a concurrent data structure, as well as its ability to adapt to situations when correctness is guaranteed by lower-level invariants (e.g., in lock-free algorithms) that are not directly expressible in the type system itself.