2020 PRIMES Spring Term Conference

Mathematics, Computer Science, and Computational Biology

June 6-7, 2020
Zoom webinar

Saturday, June 6

Mathematics Section

10:00 am Welcoming Remarks

Prof. Michel Goemans, Head of the MIT Mathematics Department
Prof. Pavel Etingof, PRIMES Chief Research Advisor
Dr. Slava Gerovitch, PRIMES Program Director

10:15 am Session 1. PRIMES STEP

- Isha Agarwal, Matvey Borodin, Aidan Duncan, Kaylee Ji, Shane Lee, Boyan Litchev, Anshul Rastogi, Garima Rastogi, and Andrew Zhao (PRIMES STEP Senior group), *A Penney for our thoughts: A story of a series of games and an Avaricious Duo* (mentor Dr. Tanya Khovanova)
- Eric Chen, William Du, Tanmay Gupta, Alicia Li, Srikar Mallajosyula, Matthew Qian, Rohith Raghavan, Arkajyoti Sinha, Maya Smith, and Samuel Wang (PRIMES STEP Junior group), *Hidden dimension of SET: The classification of magic SET squares* (mentor Dr. Tanya Khovanova)

11:00 am Session 2. PRIMES Circle

- Peter Haine, PRIMES Circle Coordinator, *Introduction*
- Anya Ditkoff and Arshia Verma, *Random numbers* (mentor Natalya Ter-Saakov)
- Kaylee Chen, Hawa Hamidou Tabayi, and Alice Zhou, *Game theory: A playful presentation* (mentor Marisa Gaetz)
- Diego González Gauss and Anthony Zhao, *Dynamic programming and applications* (mentor Jung Soo (Victor) Chu)

12:10 pm Session 3. PRIMES Circle

- Joaquín Pischner Gutierrez and Jessica He, *On group automorphisms* (mentor Zhuofan Xie)
- Zoe Shleifer and Elena Su, *Continued fractions* (mentor Maya Sankar)
- William Ayinon, *How Knot Theory is important to DNA Biology* (mentor Kenneth Cox)
- Benjamin Ratin, *Surfaces in Knot Theory* (mentor Kenneth Cox)

1:15 pm Session 4. PRIMES Circle

- Akhil Kammila and Anshul Rastogi, *Evolution of curves via curve shortening flow* (mentor Bernardo Hernández Adame)
- Divya Rajaraman and Henry Serrano-Wu, *Puzzles in graph theory* (mentor Agustin Garcia)
- Minghan Sun, Andrew Weinfeld, and Christopher Zhu (PRIMES reading group), *Homomorphisms of Graphs: Colorings, Cliques and Transitivity* (mentor Younhun Kim)

***

Sunday, June 7

Computer Science and Computational Biology Section

9:00 am Welcoming Remarks

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

9:15 am Session 1

- Sanath Govindarajan, *Integrating fully homomorphic encryption into the MLIR compiler framework* (mentor William Moses)
- Walden Yan, *Improvements on description-based neural program synthesis models* (mentor William Moses)
- Neel Bhalla, *Constructing workflow-centric traces in close to real time for the Hadoop file system* (mentor Prof. Raja Sambasivan, Tufts University)

10:20 am Session 2

- Yuxuan (Jason) Chen, *Real world application of event-based end to end autonomous driving* (mentor Dr. Igor Gilitschenski)
- Michael Gerovitch, *Visualizing and enhancing environment-aware pedestrian trajectory prediction for autonomous driving* (mentor Dr. Igor Gilitschenski)
- Ethan Mendes, *Towards a robust defense for imperceptible audio adversarial examples* (mentor Kyle Hogan)

11:25 am Session 3

- Patrick Zhang, *Privacy-preserving online advertising* (mentor Sacha Servan-Schreiber)
- Andrew Shen, *Towards verifying application isolation for cryptocurrency hardware wallets* (mentor Anish Athalye)
- Alek Westover, *The variable-processor cup game* (mentor William Kuszmaul)

12:30 pm Session 4

- Jonathan Yin, *Predicting receptor activity from structural features of chemical ligands* (mentor Dr. Hattie Chung, Broad Institute)
- Aditya Saligrama, *Revisiting ensembles in an adversarial context: Improving natural accuracy* (mentor Guillaume Leclerc)

1:10 pm: End of conference

Abstracts

Session 1. PRIMES STEP

Isha Agarwal, Matvey Borodin, Aidan Duncan, Kaylee Ji, Shane Lee, Boyan Litchev, Anshul Rastogi, Garima Rastogi, and Andrew Zhao (PRIMES STEP Senior group), "A Penney for our thoughts: A story of a series of games and an Avaricious Duo" (mentor Dr. Tanya Khovanova)

Alice and Bob attempt to resolve the allotment of a penny pile with riveting rounds of the Penney's coin game. However, upon the fateful discovery of surprising biases inherent to the game, they find themselves embarking upon another path: to create a fair coin game and return balance to their wallets. Join us alongside a plethora of definitely relevant and not at all stale puns as we explore their journey and the interesting results they unearth!

Eric Chen, William Du, Tanmay Gupta, Alicia Li, Srikar Mallajosyula, Matthew Qian, Rohith Raghavan, Arkajyoti Sinha, Maya Smith, and Samuel Wang (PRIMES STEP Junior group), "Hidden dimension of SET: The classification of magic SET squares" (mentor Dr. Tanya Khovanova)

You might have played the game of SET before, but do you know how to make a magic square out of SET cards? We classify magic SET squares using combinatorics and group theory. Last but not least, come to learn how to play a fun, new game on a magic SET square!

Session 2: PRIMES Circle

Peter Haine, PRIMES Circle Coordinator, Introduction

Anya Ditkoff and Arshia Verma, "Random numbers" (mentor Natalya Ter-Saakov)

In this presentation, we will go over the basics of discrete random variables, including expected value, variance, joint distributions, covariance, correlation, and conditional distributions. We will also go over a number of common types of distributions, including Bernoulli, Binomial, Poisson, Geometric, and Hypergeometric variables.

Kaylee Chen, Hawa Hamidou Tabayi, and Alice Zhou, "Game theory: A playful presentation" (mentor Marisa Gaetz)

In this presentation, we will explore combinatorial game theory. Combinatorial game theory is the study of games like Chess or Checkers, where two players alternate turns until one wins the game. We will explore some basic definitions and terms of game theory through two original games: Beads and Traveling. In particular, we will present a winning strategy for Beads and partial progress towards a winning strategy for Traveling.

Diego González Gauss and Anthony Zhao, "Dynamic programming and applications" (mentor Jung Soo (Victor) Chu)

In this presentation, we explore dynamic programming and its applications. Dynamic programming is a mathematical optimization method applicable in numerous fields of mathematics, from economics to bioinformatics. To give a more contextual understanding of dynamic programming we outline the concepts of time/space complexity and the hierarchy of common complexity sets. We furthermore refine the theory of dynamic programming and show examples of its use to make the idea more digestible to an audience foreign with its concepts.

Session 3: PRIMES Circle

Joaquín Pischner Gutierrez and Jessica He, "On group automorphisms" (mentor Zhuofan Xie)

In this talk, we focus on group theory and mainly explore the structure of the automorphism groups of some elementary groups. In particular, we study the automorphism groups of C_n, D_n, S_n and $(C_n)^k$ in depth. For C_n, D_n, and $(C_n)^k$, we study by tracking how the automorphisms act on the generators. For S_n, we study how most of the automorphisms are induced by conjugation and prove the theorem that only S_6 can have conjugations not induced by conjugation. At the very end, we construct one such automorphism for S_6 and calculate the total number of such automorphisms.

Zoe Shleifer and Elena Su, "π-ish: Continued fractions" (mentor Maya Sankar)

We will discuss continued fractions, an alternate representation of real numbers. We will prove that all rational numbers have a finite continued fraction representation. It turns out that continued fractions provide the best possible rational approximations of irrational numbers such as π. Finally, we will present a few additional applications of continued fractions to number theory.

William Ayinon, "How Knot Theory is important to DNA Biology" (mentor Kenneth Cox)

DNA is packaged into a cell through a process called supercoiling. When packed in the nucleus of a cell, DNA strands can become knotted. It is important to unpack the DNA so that it can interact with the organism. By using knot theory and observing how a cell unpacks its DNA, humans can better understand how living cells function.

Benjamin Ratin, "Surfaces in Knot Theory" (mentor Kenneth Cox)

Knot Theory, a branch of Topology, is a fascinating topic that studies mathematical knots. At first knots seem like a very simple concept that can be easily modeled by tying two ends of a string together. Very quickly, however, we get into complexities of knots and surfaces related to knot theory. In this presentation I will go over how we can use knot theory to better understand the properties of surfaces such as these properties include the genus, euler characteristic, and

isotopy. Using these properties we will discover that from the knot theory perspective, a sphere is mathematically the same as a cube! By the end of my presentation you will learn the basics of knot theory and you will start to see how it can be applied to the world around us!

Session 4: PRIMES Circle

Akhil Kammila and Anshul Rastogi, "Evolution of curves via curve shortening flow" (mentor Bernardo Hernández Adame)

Curve-shortening flow (CSF) is a mean geometric curvature flow with intriguing characteristics that allow it to serve as an introduction to geometric flows. In this talk we go over the basics of the flow and discuss its most interesting properties, special solutions, and applications to the fields of machine learning. In particular we discuss the Gage-Hamilton and Grayson theorems which classify the behaviour of all embedded curves under CSF.

Divya Rajaraman and Henry Serrano-Wu, "Puzzles in graph theory" (mentor Agustin Garcia)

Can you cross every bridge in a city exactly once? Can five different places be connected by non-crossing paths? In this talk, we'll introduce the basics of graph theory, and then prove results that answer these questions definitively.

Minghan Sun, Andrew Weinfeld, and Christopher Zhu (PRIMES reading group), "Homomorphisms of Graphs: Colorings, Cliques and Transitivity" (mentor Younhun Kim)

We discuss some foundational topics in algebraic graph theory as found in Godsil and Royle and some relevant results in the area. The first topic is on graph homomorphisms, colorings, and a brief discussion of Hedetniemi's Conjecture regarding the chromatic number of product graphs, which was recently disproved by Shitov. The second is about fractional cliques, fractional colorings, and their linear duality. The third is about arc-transitivity and a famous result by Weiss which states that strictly s-arc-transitive, connected regular graphs do not exist for $s > 7$.

Session 5

Sanath Govindarajan, "Integrating fully homomorphic encryption into the MLIR compiler framework" (mentor William Moses)

Fully homomorphic encryption opens up the possibility of secure computation on private data. However, fully homomorphic encryption is limited by its speed and the fact that arbitrary computations must be represented by combinations of primitive operations, such as addition, multiplication, and binary gates. Integrating FHE into the MLIR compiler infrastructure allows it to be automatically optimized at many different levels and will allow any program which compiles into MLIR to be modified to be encrypted by simply passing another flag into the compiler. The process of compiling into an intermediate representation and dynamically generating the encrypted program, rather than calling functions from a library, also allows for optimizations across multiple operations, such as rewriting a DAG of operations to run faster and removing unnecessary operations.

Walden Yan, "Improvements on description-based neural program synthesis models" (mentor William Moses)

With the recent rise in interest in artificial neural networks and deep learning, the new AI paradigm has been adapted to the task of neural program synthesis (NPS). In NPS, a model is designed to output code that solves a given program specification. However, most existing models rely on too complex an underlying structure that makes effective training difficult. Linear models, however, are hindered by the fact that relationships between tokens are hard to identify. In this work, we take a simple linear model and modify it to provide greater structure and streamline the reproduction of common subprocedures. We focus on natural language specifications as opposed to input/output (I/O) examples. Not only are natural language descriptions easier to obtain, but they also expand our domain to more complex programs that would be impossible to infer from just I/O pairs. Our model is trained and evaluated on AlgoLisp, a dataset of problem descriptions paired with example solutions and test cases with which to evaluate programs. Borrowing the most successful techniques from previous works, we use a neural network to guide a beam search. The architecture consists of an encoder and decoder LSTM augmented with attention mechanisms and a specialized syntax layer. Motivated by the technique of byte-pair encoding (BPE) prevalent in natural language processing (NLP) models, we also feed our model an expanded vocabulary of AlgoLisp's domain-specific language (DSL) generated by pairing commonly adjacent tokens. In the end, our model achieves 98.9% accuracy at evaluation, which greatly improves on previous state-of-the-art (95.8%), while using a comparable number of parameters.

Neel Bhalla, "Constructing workflow-centric traces in close to real time for the Hadoop file system" (mentor Prof. Raja Sambasivan, Tufts University)

Diagnosing problems in large-scale distributed services is like finding a needle in a haystack. Such services can be comprised of 1000s of individual nodes. Any subset of these nodes could be involved in a given request's processing and responsible for observed problems. To help,

recent work on workflow-centric tracing records the order and timing of requests' execution within and among distributed-service nodes (i.e., records their workflows). But, existing tracing systems are unable to make traces available soon after requests' execution, limiting their applicability. In this work, we demonstrate how real-time stream processing systems can be modified and used to construct traces in real time. We also demonstrate how such real-time trace construction can be used for anomaly detection. The tracing system is currently being integrated into Pythia, a system for tuning instrumentation for traces, as a feedback mechanism for determining the granularity of traces.

Session 6

Yuxuan (Jason) Chen, "Real world application of event-based end to end autonomous driving" (mentor Dr. Igor Gilitschenski)

End-to-end autonomous driving has recently been a popular area of study for deep learning. This work studies the use of event cameras for real-world deep learned driving task in comparison to traditions RGB cameras. In this work, we evaluate existing state-of-the-art event-based models on offline datasets, design a novel model that fuses the benefits from both event-based and traditional frame-based cameras, and integrate the trained models on board a full-scale vehicle. We conduct tests in a challenging track with features unseen to the model. Through our experiments and saliency visualization, we show that event-based models actually predict the existing motion of the car rather than the active control the car should take. Therefore, while event-based models excel at offline tasks such as motion estimation, our experiments reveal a fundamental challenge in applying event-based end-to-end learning to active control tasks, that the models need to learn reasoning about future actions with a feedback loop that impacts its future state.

Michael Gerovitch, "Visualizing and enhancing environment-aware pedestrian trajectory prediction for autonomous driving" (mentor Dr. Igor Gilitschenski)

We apply the deep learning approach to predict future trajectories of dynamic objects, such as pedestrians crossing a street, to aid autonomous driving. We create a convolutional neural network to generate future trajectories of the pedestrians and train this network, using pre-annotated video footage of a crowded square and a subway station. We generate an interchangeable location bias map to account for changes in scenery, including immovable objects, and incorporate it into the network. To speed up training, we initially train the network without the location bias map, and then add the bias map and adjust the hyperparameters. In addition, we implement a visualization component that overlays input and output vectors onto the video footage to gain a better understanding of the accuracy of our trajectory predictor.

Ethan Mendes, "Towards a robust defense for imperceptible audio adversarial examples" (mentor Kyle Hogan)

Neural networks are susceptible to adversarial examples, which are specific inputs to a network that result in a misclassification or an incorrect output. While most past work has focused on methods to generate adversarial examples to fool image classification networks, recently, similar attacks on automatic speech recognition systems have been explored. Due to the relative novelty of these audio adversarial examples, there exist few robust defenses for these attacks. In this talk, I present a robust defense for inaudible or imperceptible audio adversarial examples. This approach mimics the adversarial strategy to add targetted proportional additive gaussian noise in order to revert an adversarial example back to its original transcription. Additionally, I demonstrate the challenges that arise when applying defenses against adversarial examples for images to audio adversarial examples.

Session 7

Patrick Zhang, "Privacy-preserving online advertising" (mentor Kyle Hogan)

Online advertisements are an essential component of the business model of many online platforms and often the main source of revenue. Ads are most effective when relevant ads are shown to the proper audience, which requires detailed personal information to be collected on web users. This is accomplished using Similarity Search algorithms which match user interests to ads. However, current techniques make no attempt at providing privacy to the end users. We propose a novel approach to the similarity search problem for selecting ads by designing a private interactive protocol that provides user privacy without compromising ad selection accuracy. Our protocol matches the accuracy of existing Similarity Search algorithms and is experimentally evaluated using real-world data.

Andrew Shen, "Towards verifying application isolation for cryptocurrency hardware wallets" (mentor Anish Athalye)

We often perform security-sensitive operations in our day-to-day lives such as performing monetary transactions. To perform these operations securely, we can isolate the confirmation of such operations to separate hardware devices. However, proving that these devices operate securely is still difficult given the complexity of their kernels, yet important given the rise in popularity of cryptocurrency transaction devices. To support multiple cryptocurrencies and other functionality, these devices must be able to run multiple applications that are isolated from one another as they could be potentially maliciously acting applications. We can simplify our device by modeling it as running applications sequentially in user mode. We seek to prove that these applications cannot tamper with the kernel memory and show that the kernel protection is set up correctly. To do this, we developed a RISC-V machine emulator in Rosette, which enables us to reason about the behaviour of symbolic machine states and symbolic applications. We make progress towards verifying application isolation for launching and running applications on a simple kernel.

Alek Westover, "The variable-processor cup game" (mentor William Kuszmaul)

In a *cup game* two players, the *filler* and the *emptier*, take turns adding and removing water from cups, subject to certain constraints. In the classic p-processor cup game the filler distributes p units of water among the n cups with at most 1 unit of water to any particular cup, and the emptier chooses p cups to remove at most one unit of water from. Analysis of the cup game is important for applications in processor scheduling, buffer management in networks, quality of service guarantees, and deamortization.

We investigate a new variant of the classic p-processor cup game, which we call the *variable-processor cup game*, in which the resources of the emptier and filler are variable. In particular, in the variable-processor cup game the filler is allowed to change p at the beginning of each round. Although the modification to allow variable resources seems small, we show that it drastically alters the game.

We construct an adaptive filling strategy that achieves backlog $Omega(n^{1-\epsilon})$ for any constant $\epsilon>0$ of our choice in running time $2^{O(\log^2 n)}$. This is enormous compared to the upper bound of $O(\log n)$ that holds in the classic p-processor cup game!

We also present a simple adaptive filling strategy that is able to achieve backlog $Omega(n)$ in extremely long games: it has running time $2^{O(n)}$.

Furthermore, we demonstrate that this lower bound on backlog is tight: using a novel set of invariants we prove that a greedy emptier never lets backlog exceed $O(n)$.

We also construct an oblivious filling strategy that achieves backlog $Omega(n^{1-\epsilon})$ for $\epsilon>0$ constant of our choice in time $2^{O(\log^2 n)}$ against any ``greedy-like" emptier with probability at least $1-2^{-polylog(n)}$. Whereas classically randomization gives the emptier a large advantage, in the variable-processor cup game the lower bound is the same!

Session 8

Jonathan Yin, "Predicting receptor activity from structural features of chemical ligands" (mentor Dr. Hattie Chung, Broad Institute)

Receptors on cellular surfaces detect and compute environmental signals. In the olfactory system, millions of small molecule ligands are sensed by hundreds of olfactory receptors (ORs) through combinatorial encoding. The ability to predict activated receptors based on the chemical structure of ligands is a powerful tool with applications in drug design. While traditional machine learning models typically perform well when trained on large, high-quality datasets, obtaining such data in the biological realm remains challenging due to experimental and time constraints. Here we aim to work around these data constraints by implementing a two-step process that first extracts and encodes salient chemicals features and then uses that latent representation as input to train a prediction model. While existing abstraction models rely heavily on lower-level syntax-based methods, we incorporate meaningful higher-level molecular features through functional group encodings alongside traditional approaches. Using this improved continuous molecular representation, we strive to reduce the data needed to train a prediction model and better understand the responsible molecular features.

Aditya Saligrama, "Revisiting ensembles in an adversarial context: Improving natural accuracy" (mentor Guillaume Leclerc)

A necessary characteristic for the deployment of deep learning models in real world applications is resistance to small adversarial perturbations while maintaining accuracy on non-malicious inputs. While robust training provides models that exhibit better adversarial accuracy than standard models, there is still a significant gap in natural accuracy between robust and non-robust models which we aim to bridge. We consider a number of ensemble methods designed to mitigate this performance difference. Our key insight is that model trained to withstand small attacks, when ensembled, can often withstand significantly larger attacks, and this concept can in turn be leveraged to optimize natural accuracy. We consider two schemes, one that combines predictions from several randomly initialized robust models, and the other that fuses features from robust and standard models.