MIT Stata Center. Photo: Slava Gerovitch

Eleventh Annual Spring-Term PRIMES Conference

May 22-23, 2021

2021 PRIMES Spring Term Conference

Saturday, May 22

Mathematics

9:00 am: Welcoming Remarks

Prof. Pavel Etingof, PRIMES Chief Research Advisor
Dr. Slava Gerovitch, PRIMES Program Director

9:15-10:15 am: Session 1: PRIMES Circle

Gloria Chun & Alicia Li, “The Post Correspondence Problem” (mentor Alexandra Hoey)

Isha Agarwal & Minseo Kim, “DP-3T: Ensuring Cryptographic Safety and Security In COVID-19 Contact Tracing” (mentor Talia Blum)

Kien Lau & Andrew Zeng, “Models from Introductory Probability" (mentor Jordan Benson)

10:20-11:20 am: Session 2: PRIMES Circle

Lisa Liu & Tanvi Ganapathy, “A Brief Review of Special Topics in Group Theory” (mentor Gabrielle Kaili-May Liu)

Xavier Choe & Garima Rastogi, “Number Fields and Galois Theory” (mentor Merrick Cai)

Dimitar Dimitrov & Isaac Patterson, “Invariants in Knot Theory” (mentor Preston Cranford)

11:25 am - 12:20 pm: Session 3: PRIMES Circle

Neil Jin & Apurva Varigonda, “Graph Theory and the Optimization of City Infrastructure” (mentor Rachana Madhukara)

Anusha Senapati & Iris Heikkinen, “Branches of Graph Theory: On Spanning Trees” (mentor Rebecca Nelson)

Ephram Chun, “Applications of Combinatorial Algorithms for Graph Traversal and Efficiency” (mentor Jung Soo Chu)

12:30-1:00 pm: Virtual Chat with PRIMES Chief Research Advisor Prof. Pavel Etingof, Head Mentor Dr. Tanya Khovanova, Circle coordinator Peter Haine, and assistant Circle coordinator Marisa Gaetz (for students and parents)

1:15-2:10 pm: Session 4: PRIMES Circle

Christian Nichols & Duy Pham, “Combinatorics” (mentor Eli Garcia)

Shreya Sinha & Jennifer Yuan, “Well, that's just a theory.... a GAME theory” (mentor Katie Gravel)

Saniya Srivastava & Heidi Zhang, “Modeling Altruism in Evolutionary Biology Using Game Theory” (mentor Ariana Park)

2:15-3:10 pm: Session 5: PRIMES Reading and PRIMES STEP

Jason Tang (PRIMES Reading), “The Sperner Property and Quotients of Boolean Algebras” (mentor Luis Kumanduri)

Michael Han, Ella Kim, Evin Liang, Mira Lubashev, Oleg Polin, Vaibhav Rastogi, Benjamin Taycher, Ada Tsui, Cindy Wei (PRIMES STEP Junior group), “Fun with Latin Squares” (mentor Dr. Tanya Khovanova)

Matvey Borodin, Eric Chen, Aidan Duncan, Boyan Litchev, Jiahe Liu, Veronika Moroz, Matthew Qian, Rohith Raghavan, Garima Rastogi, and Michael Voigt (PRIMES STEP Senior group), “Stable Marriage Problem” (mentor Dr. Tanya Khovanova)

Sunday, May 23

Computer Science and Computational Biology

9:00 am Welcoming Remarks

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

9:10-10:10 am: Session 6: Computer Science

Benjamin Chen, “Practical Anonymity Sets in Pseudonymous Forums” (mentor Kyle Hogan)

Linda Chen, “Communication Complexity of Byzantine Broadcast” (mentor Jun Wan)

Matthew Ding, “Signed-Message Relay Version of Iterative Approximate Byzantine Consensus for Directed Graphs” (mentor Hanshen Xiao)

10:20-11:20 am: Session 7: Computer Science and PRIMES Reading

Beining (Cathy) Zhou, “A High-Order Cumulant-Based Sparse Ruler for Improved Lag Generation” (mentor Hanshen Xiao)

Jason Yang, “Decentralized Gradient Descent: How Network Structure Affects Convergence” (mentors Jun Wan and Hanshen Xiao)

Ishita Goluguri and Toyesh Jayaswal (PRIMES Reading), “Jumping into Markov Chains” (mentor Dr. Benjamin Landon)

11:30 am - 12:10 pm: Session 8: Computational Biology

Sarah Chen, “*In Silico* Prediction of Retained Intron-Derived Neoantigens in Leukemia” (mentors Dr. Nicoletta Cieri, Dana-Farber Cancer Institute, and Kari Stromhaug)

Mikhail Alperovich, “Data Driven Quality Control for Single-Cell RNA Sequencing” (mentor Dr. Ayshwarya Subramanian, Broad Institute)

2021 PRIMES CONFERENCE ABSTRACTS

SATURDAY, MAY 22

SESSION 1: PRIMES CIRCLE

Gloria Chun & Alicia Li

The Post Correspondence Problem

Mentor: Alexandra Hoey

The Post Correspondence Problem is a decision problem that takes as input a sequence of dominos

$$\left\{\left[\frac{t_1}{b_1}\right], \left[\frac{t_2}{b_2}\right], \dots, \left[\frac{t_n}{b_n}\right]\right\}$$

and returns whether there exists a match, i.e., a sequence $i_1, i_2, \dots, i_\ell$ such that $t_{i_1} t_{i_2} \cdots t_{i_\ell} = b_{i_1} b_{i_2} \cdots b_{i_\ell}$. In this talk, we will explain basic concepts from computability theory such as Turing machines and reductions, and by presenting a reduction from the undecidable language A_{TM}, we will show that there is no programmable algorithm that can solve the Post Correspondence Problem.

Isha Agarwal & Minseo Kim

DP-3T: Ensuring Cryptographic Security and Privacy In COVID-19 Contact Tracing

Mentor: Talia Blum

During the COVID-19 pandemic, people have been taking various social as well as innovative measures in order to combat the virus. One of these solutions is contact tracing, which encompasses many different methods, both centralized and decentralized, to identify people who might have been exposed to COVID-19. In our presentation, we deconstruct the mathematics behind the Decentralized Privacy-Preserving Proximity Tracing (DP-3T) algorithm for digital contact tracing and determine some of its potential vulnerabilities. After an analysis of DP-3T, we will provide a brief synopsis of our own proposed improvements, drawing from both pre-existing and novel systems—utilizing RSA encryption and the BB84 quantum computing algorithm—to enhance the cryptographical security of DP-3T.

Kien Lau & Andrew Zeng

Models from Introductory Probability

Mentor: Jordan Benson

This paper will cover the introductory ideas of counting principles that will lead to the fundamentals of probability as sets, exploring operations in set theory such as unions and intersections.Using these ideas, we will apply them to the Statistical field, summarizing the Binomial and Poisson distributions respectively. We show that a simple binomial model can predict weather patterns to unexpected accuracy!

SESSION 2: PRIMES CIRCLE

Tanvi Ganapathy & Lisa Liu

A Brief Review of Special Topics in Group Theory

Mentor: Gabrielle Kaili-May Liu

In this presentation, we introduce groups and group theory via a series of key definitions and theorems, and build upon this foundation to discuss several special topics in group theory—namely, *wallpaper groups* and *Tarski groups*. Wallpaper groups are symmetry groups in two dimensions, which we will introduce by first exploring C_n, D_n, and Frieze Groups. Tarski monsters are infinite groups whose non-trivial proper subgroups are all of the same prime order p.

Xavier Choe & Garima Rastogi

Number Fields and Galois Theory

Mentor: Merrick Cai

In this presentation, we will give an overview of number fields and Galois theory applied to number fields. We focus on the construction of number fields via the primitive element theorem, and using this approach, we can understand how certain ideals factor into prime ideals. We also give an overview of Galois theory and discuss the Galois correspondence, understanding subgroups of the relative automorphism groups and their relation to intermediate fields.

Dimitar Dimitrov & Isaac Patterson

Invariants in Knot Theory

Mentor: Preston Cranford

We introduce the basics of knot theory then discuss several invariants appearing in knot theory including linking number, tricolorability, the bracket polynomial, and the Jones polynomial.

SESSION 3: PRIMES CIRCLE

Neil Jin & Apurva Varigonda

Graph Theory and the Optimization of City Infrastructure

Mentor: Rachana Madhukara

An introduction to graph theory is presented alongside an application of this theory to the optimization of city infrastructure. In particular, the presentation is split into three sections. First, the relevant definitions of graphs and their properties are explained along with examples. Then, some interesting properties of graphs and maps between graphs are explored. In this context, some preliminary theorems are also proven. Lastly, the notion of an Eulerian graph is introduced and its properties are explained through the proof of a famous theorem. The presentation is concluded with an explanation of how Eulerian graphs can be used to optimize the path of delivery agents through a city.

Anusha Senapati & Iris Heikkinen

Branches of Graph Theory: On Spanning Trees

Mentor: Rebecca Nelson

The purpose of our presentation is to introduce our peers to the basic concepts of graph theory and later examine the applications of these concepts. This includes going over terms such as nodes, edges, size, degree, adjacent matrices, etc. In addition, we will introduce a new idea: spanning trees. We plan to go over the Minimum Spanning Tree Problem from *A First Course in Graph Theory* by Gary Chartrand and Ping Zhang and then go over a few related practice problems. From here, we will be able to show the applications of the Spanning Tree Graphs and notice how it helps calculate the shortest path possible (to connect all vertices of a graph).

Ephram Chun

Applications of Combinatorial Algorithms for Graph Traversal and Efficiency

Mentor: Jung Soo (Victor) Chu

An algorithm is a finite sequence of unambiguously defined steps that carries out a task. In this presentation, we will explore various topics in combinatorics, including proof techniques, graph theory, game theory, and algorithms. Lastly, we will apply these topics to present and explore a solution to a problem where we assign classes to classrooms of a high school while minimizing the students' travel distances between classes.

SESSION 4: PRIMES CIRCLE

Christian Nichols & Duy Pham

The binomial theorem and related identities

Mentor: Elias Garcia

In this talk, we will discuss some combinatorial identities, such as the binomial theorem, relations between binomial coefficients, and Vandermonde's theorem.

Shreya Sinha & Jennifer Yuan

Well, that's just a theory…a GAME theory

Mentor: Katherine Gravel

Game theory is the study of interactive and strategic decision-making among people in a group or game where one's strategy depends on the choices of other players. It has applications in all the social sciences (most notably economics), systems science, logic, as well as computer science. In this presentation, we will discuss one subclass of games studied in game theory, called combinatorial games. We will also explore a particular combinatorial game called Nim and understand some of its surprising properties that can be extended to other combinatorial games.

Saniya Srivastava & Heidi Zhang

Modeling Altruism in Evolutionary Biology Using Game Theory

Mentor: Ariana Park

In this presentation, we model altruism in vampire bats using three different strategies based in classical game theory. Vampire bats demonstrate altruism by regurgitating blood to give to other vampire bats in order to prevent them from dying of hunger. First, we use a simple 2-player matrix game and a utility function to observe how vampire bats value the health of a group over their own health. Building on our simple model, we use a CC-PP game model and a repeated game model to model more nuanced factors that affect how and when a vampire bat chooses to regurgitate blood.

Jason Tang (PRIMES Reading)

The Sperner Property and Quotients of Boolean Algebras

Mentor: Luis Kumanduri

We explore partially-ordered sets that satisfy the Sperner Property, a condition that bounds the size of a poset's antichains in relation to its levels. In particular, we prove that both the Boolean algebra itself and its quotient posets are Sperner, introducing order-matchings and order-raising operators along the way. We finish with applications of these results to various common combinatorial objects.

Michael Han, Ella Kim, Evin Liang, Mira Lubashev, Oleg Polin, Vaibhav Rastogi, Benjamin Taycher, Ada Tsui, Cindy Wei (PRIMES STEP Junior group)

Fun with Latin Squares

Mentor: Dr. Tanya Khovanova

Do you want to know what an anti-chiece Latin square is? Or what a non-consecutive toroidal modular Latin square is? Come to our talk! We invented a ton of new types of Latin squares, some inspired by existing Sudoku variations, and we created some new words to describe our Latin squares. We can't wait to introduce them to you and answer other questions, like do they even exist? If so, under what conditions? What are some of their interesting properties? And how do we generate them?

Matvey Borodin, Eric Chen, Aidan Duncan, Boyan Litchev, Jiahe Liu, Veronika Moroz, Matthew Qian, Rohith Raghavan, Garima Rastogi, and Michael Voigt (PRIMES STEP Senior group)

Stable Marriage Problem

Mentor: Dr. Tanya Khovanova

Are you having trouble getting married? These days there are lots of products on the market for dating, from apps to websites and matchmakers, but we know a simpler way! That's right, your path to coupled life isn't through Tinder, it's through Sudoku! Come to our fabulous talk where we explore the Stable Marriage Problem to help you find happiness and stability in marriage and math.

SESSION 6: COMPUTER SCIENCE

Benjamin Chen

Practical Anonymity Sets in Pseudonymous Forums

Mentor: Kyle Hogan

Pseudonymous forums are online websites where users can post publicly visible content and participate in discussions under a pseudonym. Such forums are not perfectly private, as their privacy can be compromised to traffic analysis attacks. However, many methods of providing perfect privacy to such a system come with a heavy performance cost—whether in bandwidth or latency. We examine the practicality of anonymity sets, a defense against such attacks that can still provide a formal privacy guarantee with less performance losses, and attempt to simulate their implementation in a real-world setting using real data scraped from Reddit, a popular pseudonymous forum. We try various methods of creating these anonymity sets, finding that K-means with some dimensionality compression yields decent results; we also propose a method of defining a common traffic budget for members of a set. We find that anonymity sets are a feasible defense against such attacks, and we also discuss what information is leaked by such sets.

Linda Chen

Communication Complexity of Byzantine Broadcast

Mentor: Jun Wan

Byzantine Broadcast is a fundamental problem in distributed computing, with communication complexity being an important aspect of Byzantine Broadcast protocols. In Byzantine Broadcast, a designated leader must ensure that all honest users in a distributed system reach a consensus, even in the presence of some dishonest users. Previous works have shown an $O(n^2)$ lower bound on communication complexity, as well as protocols with $O(n^2)$ communication complexity for the honest majority scenario. In this project, we provide various methods and intuition towards a possible $O(n^3)$ communication complexity lower bound for dishonest majority Byzantine Broadcast.

Matthew Ding

Signed-Message Relay Version of Iterative Approximate Byzantine Consensus for Directed Graphs

Mentor: Hanshen Xiao

This paper presents a novel algorithm for Iterative Approximate Byzantine Consensus, called Relay-IABC. The algorithm allows machines to achieve approximate consensus to arbitrary exactness in the presence of byzantine failures. The algorithm relies on the novel usage of a relayed messaging system and signed messages with unforgeable signatures that are unique to each node. The use of signatures and relays allows the strict necessary network conditions of traditional approximate byzantine consensus algorithms to be circumvented. In addition, we show that the relay system achieves faster convergence even with the strict network conditions. Previous methods which use transition matrices to prove convergence are extended by having each state vector model not just one iteration, but a set of D iterations, where D is a diameter property of the graph. This allows us to accurately model the messages within the relay system.

Beining (Cathy) Zhou

A High-Order Cumulant-Based Sparse Ruler for Improved Lag Generation

Mentor: Hanshen Xiao

The sparse ruler is a classical problem in theoretical computer science that seeks to find a sequence of integers, or integer marks on a ruler, such that the differences of the marks can generate as many consecutive integers lag as possible. Most recently, the sparse ruler is generalized with high-order cumulants to optimize the lags generated. We will present a new construction of the ruler for the 4-th and 6-th order and extend it to the $2q$-th order by layering. Compared to two other prominent methods, this approach significantly improves the lag generation by exploiting more sign combinations of the permutation invariants. This problem and the construction could also be applied to spatial signal processing to provide more efficient sensor array geometries.

Jason Yang

Decentralized Gradient Descent: How Network Structure Affects Convergence

Mentors: Jun Wan and Hanshen Xiao

Machine learning is becoming increasingly decentralized and reliant on the cooperation of multiple agents in a network. Because of this, we investigate decentralized gradient descent over such a network, where certain agents are corrupted by an adversary. We focus on the case where the functions of all nodes are 1-dimensional quadratics, and where each corrupted agent is connected to all honest agents.

Ishita Goluguri and Toyesh Jayaswal (PRIMES Reading)

Jumping into Markov Chains

Mentor: Dr. Benjamin Landon

In this presentation, we discuss Markov Chains and some of their applications through examples. A Markov Chain is a process where the $n+1$th step only depends on the nth step and is independent of the previous steps. After the introduction, we discuss hitting times and probabilities, and work through an example called the Gamblers' Ruin. Then, we define recurrence and transience, which describe long-term behavior of the Markov Chain, and explore a few examples of random walks on integer lattices. Finally, we study the invariant distribution of the Markov Chain and the Ergodic theorem, which describes the long-term probability distribution of where the Markov Chain is.

SESSION 8: COMPUTATIONAL BIOLOGY

Sarah Chen

In Silico *Prediction of Retained Intron-Derived Neoantigens in Leukemia*

Mentor: Dr. Nicoletta Cieri, Dana-Farber Cancer Institute, and Kari Stromhaug

Alternative splicing is critical for the regulation and diversification of gene expression. Conversely, splicing dysregulation, caused by mutations in splicing machinery or splice junctions, is a hallmark of cancer. Tumor-specific isoforms are a potential source of neoantigens, cancer-specific peptides presented by human leukocyte antigen (HLA) class I molecules and potentially recognized by T cells. For cancers such as acute myeloid leukemia (AML) with a low mutation burden but widespread splicing aberrations, splice variants and retained introns (RIs) in particular, may broaden the number of suitable targets for immunotherapy. I developed a computational pipeline to predict AS-derived neoepitopes

from tumor RNA-Seq. I first used the B721.221 B cell line as a model system, for which RNA-Seq, Ribo-Seq, and immunoproteome data from >90 HLA class I monoallelic lines were available. I performed de novo transcriptome assembly with StringTie, identifying on average 694±73 AS isoforms across 4 technical replicates. Using HLAthena, I identified 1,087 AS-derived neoepitopes predicted to bind across 4 frequent HLA alleles. Of them, 192 (18%) also displayed evidence of mRNA translation, measured as the alignment of ≥1 Ribo-Seq. To further increase prediction accuracy, I am currently analyzing the HLA I immunopeptidome to define the features of predicted AS isoforms more likely to be not only translated but also HLA presented. Finally, I applied my prediction pipeline to AML cell lines (n=8) and primary samples (n=7). I identified 682±113 AS isoforms in AML cell lines, similar to the 694 in B721, but the proportion of isoforms containing RIs (as opposed to alternative 5' and 3' splice sites or cassette exons) was 3.5x higher than in B721, in line with the biological relevance of RIs in particular in this disease setting. Primary AML samples yielded 1496±294 AS isoforms, more than twofold the number in B721 or AML cell lines, thus reinforcing the significant contribution of AS to the cancer immunopeptidome. Accurate prediction of AS-derived neoantigens through this pipeline will contribute to the design of novel cancer immunotherapies.

Mikhail Alperovich

Data Driven Quality Control for Single-Cell RNA Sequencing

Mentor: Dr. Ayshwarya Subramanian, Broad Institute

Quality control (QC) of cells, a critical step in single-cell RNA sequencing data analysis, has largely relied on arbitrarily fixed data-agnostic thresholds on QC metrics such as gene complexity and fraction of reads mapping to mitochondrial genes. The few existing data-driven approaches perform QC at the level of samples or studies without accounting for biological variation in the commonly used QC criteria. We demonstrate that the QC metrics vary both at the tissue and cell state level across technologies, study conditions, and species. We propose ddqc, an unsupervised adaptive quality control framework that performs flexible and data-driven quality control at the level of cell states while retaining critical biological insights and improved power for downstream analysis. On applying ddqc to 6,228,212 cells and 835 mouse and human samples, we retain a median of 39.7% more cells when compared to conventional data-agnostic QC filters. With ddqc, we recover biologically meaningful trends in gene complexity and ribosomal expression among cell types enabling exploration of cell states with minimal transcriptional diversity or maximum translational potential. Moreover, ddqc allows us to retain cell types often lost by conventional QC such as metabolically active parenchymal cells, and specialized cells such as neutrophils or gastric chief cells. Taken together, our work leads us to propose a revised paradigm to quality filtering best practices - iterative QC, providing a data-driven quality control framework compatible with observed biological diversity.