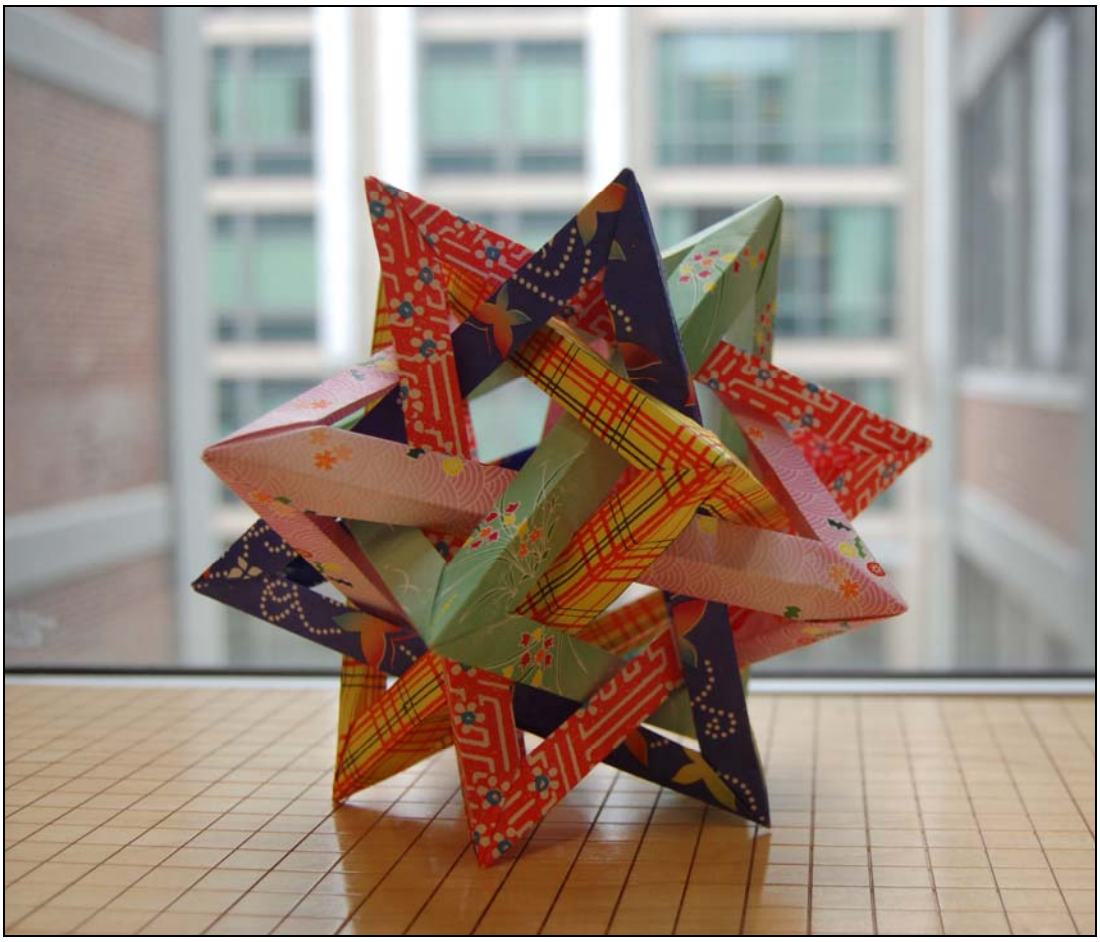
MIT PRIMES



Five tetrahedra in the regular dodecahedron. By Sheela Devadas, PRIMES student in 2011-12 and the winner of the 2015 Alice T. Schafer Prize for excellence in mathematics by an undergraduate woman. Photo by Dennis Porche.

FIFTH ANNUAL CONFERENCE

MAY 2015

2015 MIT PRIMES CONFERENCE

Program for Research In Mathematics, Engineering, and Science for High School Students

Five tetrahedra in the regular dodecahedron. By Sheela Devadas, PRIMES student in 2011-12 and the winner of the 2015 Alice T. Schafer Prize for excellence in mathematics by an undergraduate woman. Photo by Dennis Porche.

Saturday, May 16: Mathematics

8:30 am Welcoming remarks
Prof. Tomasz Mrowka, Head of the MIT Mathematics Department
Prof. Pavel Etingof, PRIMES Chief Research Advisor
Dr. Slava Gerovitch, PRIMES Program Director

9:00 am Session 1
Varun Jain, *Circular planar graphs and electrical networks* (mentor Carl Lian)
Alok Puranik, *Limitations of semidefinite programming for certifying RIP* (mentor Adrian Vladu)
Kavish Gandhi and Noah Golowich, *Analysis of Boolean functions* (mentor Yufei Zhao)

10:25 am Session 2
Dhruv Medarametla, *Bounds on the norms of locally random matrices* (mentor Aaron Potechin)
Karan Sarkar, *On modular extensions to Nim* (mentor Dr. Tanya Khovanova)
Caleb Ji, Robin Park, and Angela Song, *Combinatorial games of no strategy* (mentor Dr. Tanya Khovanova)

11:45 am Session 3: PRIMES-IGL
Mehtaab Sawhney, *A study of bar and arc k-visibility graphs* (mentor Jonathan Weed)
Richard Yi, *Continuous model for two-lane traffic flow* (mentor Prof. Gabriele LaNave, University of Illinois at Urbana-Champaign)
Daniel Guo, *An infection spreading model on trees* (mentor Prof. Partha Dey, University of Illinois at Urbana-Champaign)

1:45 pm Session 4
Girishvar Venkat, *Signatures of the contravariant form on Specht modules for cyclotomic Hecke algebras* (mentor Siddharth Venkatesh)
Samuel Rush, *Signatures in representations of rational Cherednik algebras* (mentor Gus Lonergan)
Luke Sciarappa, *Algebras in representations of the symmetric group S_t, when t is transcendental* (mentor Nate Harman)

3:00 pm Session 5
Brandon Epstein, *The defect angle and the relation to the Laplacian matrix* (mentor Prof. Martin Rocek, SUNY at Stony Brook)
Rachel Zhang, *Statistics of intersections of curves on surfaces* (mentor Prof. Moira Chas, SUNY at Stony Brook)
Arthur Azvolinsky, *Explicit computations of the frozen boundaries of rhombus tilings* (mentor Alisa Knizel)

4:15 pm Session 6
Meena Jagadeesan, *The exchange graphs of maximal weakly separated collections* (mentor Miriam Farber)
Meghal Gupta, *Extremal functions of forbidden matrices* (mentor Jesse Geneson)
David Amirault, *Better bounds on the rate of non-witnesses of Lucas pseudoprimes* (mentor David Corwin)

5:25 pm Session 7
Jacob Klegar, *Tiling-harmonic functions* (mentor Prof. Sergiy Merenkov, CCNY-CUNY)
Ahaan Rungta, *Mathematically modeling the motion of cells in porous media* (mentor Andrew Rzeznik)
Nick Diaco, *A new coin weighing problem and concealing information* (mentor Dr. Tanya Khovanova)

Sunday, May 17: Computer Science and Computational Biology

8:30 am Welcoming remarks
Prof. Srini Devadas, MIT Department of Electrical Engineering and Computer Science
Dr. Slava Gerovitch, PRIMES Program Director

8:45 am Session 8: Medical Informatics
Ashay Athalye, *Machine learning characterization and prediction of intrinsically disordered protein interactions: A focus on BRCA1* (mentor Dr. Gil Alterovitz)
Arul Prasad, *The significance of disordered residues in bacterial drug resistance and SNP interactions in relation to disease associations* (mentor Dr. Gil Alterovitz)
Kara Luo, *Computer simulation of biosynthetic drug modifications to improve binding activity* (mentor Dr. Gil Alterovitz)

10:00 am Session 9: Medical Informatics and Computational Biology
Andrew Li, *Exploring multi-conformational modeling and flexibility of molecular recognition features in improving drug docking* (mentor Dr. Gil Alterovitz)
Daniel Lu, *Investigating drug synergy mechanisms of disordered protein-related diseases* (mentor Dr. Gil Alterovitz)
Laura Braverman and Betsy Pu, *Genomic and epigenomic signatures of chromosomal domains* (mentors Maxim Imakaev and Boryana Doyle)

11:15 pm Session 10: Computer Science
Amy Chou and Justin Kaashoek, *Automating generation of programming problems* (mentor Rohit Singh)
Harshal Sheth and Aashish Welling, *A garbage collected network stack with CSP threads* (mentor Cody Cutler)
Gregory Barboy, Albert Gerovitch, and Andrew Gritsevskiy, *Mobile health surveillance: The development of software tools for monitoring the spread of disease* (mentor Dr. Natasha Markuzon, Draper Lab)

1:30 pm Session 11: Computational Biology
Michael Colavita, *Clustering of pathogenic genes in human coregulatory network* (mentor Soheil Feizi)
Allison Paul, *The inference of directed acyclic graphs using spectral clustering* (mentor Soheil Feizi)
Lalita Devadas, *Modelling changes in gene expression using five histone modifications* (mentor Angela Yen)

2:45 pm Session 12: Computer Science
Diana Ding and Cristian Gutu, *SecretRoom: An anonymous chat client* (mentor Albert Kwon)
Akiva Gordon and Krishna Suraj, *Improving oblivious RAM protocol through novel eviction and access strategies* (mentor Ling Ren)
Henry Liu and Ethan Zou, *Time traveling in multicore processors* (ment

4:15 pm Session 13: Mathematics
Uma Roy, *Infinity crystals for certain generalized quantum groups* (mentor Seth Shelley-Abrahamson)
Niket Gowravaram, *XYX Algebras* (mentor Dr. Tanya Khovanova)
Eric Nie, *Dual Schubert polynomials* (mentor Pavel Galashin)

5:25 pm Session 14: Mathematics
Kenz Kallal, Matt Lipman, and Felix Wang, *Equal compositions of rational functions* (mentor Thao Thi Thu Do and Prof. Michael Zieve)
Arjun Khandelwal and Joshua Xiong, *Linear algebra methods in combinatorics* (mentor Chiheon Kim)

MIT MATHEMATICS | MIT Physical Sciences - Oncology Center | EECS | DRAPER LABORATORY
MathWorks | Quanta Computer | National Science Foundation WHERE DISCOVERIES BEGIN | CMI
Gabriella & Paul Rosenbaum Foundation | SIMONS FOUNDATION

Room 4-370, MIT
web.mit.edu/primes

2015 PRIMES CONFERENCE ABSTRACTS

SATURDAY, MAY 16
MATHEMATICS

SESSION 1

Varun Jain

Circular planar graphs and electrical networks

Mentor: Carl Lian

Project suggested by Carl Lian

Circular planar graphs are used to model electrical networks, which arise in classical physics. Based on network response matrices, these circular planar graphs can be organized into equivalence classes. We first investigate the number of Y-Delta transformations required to transform one critical graph in an equivalence class to another. Next, we consider positivity phenomena, studying how testing the signs of certain circular minors can be used to determine if a given network response matrix is associated with a particular equivalence class. In particular, we prove a conjecture by Kenyon and Wilson for a few cases.

Alok Puranik

Limitations of semidefinite programming for certifying RIP

Mentor: Adrian Vladu

Project suggested by Adrian Vladu

The Restricted Isometry Property (RIP) characterizes matrices that approximately preserve the norms of sparse vectors. Introduced by Candès and Tao, it is an instrumental concept in compressive sensing. Most previous work on RIP concerns construction of such matrices under various parameter regimes, but we consider the computational problem of certifying whether a matrix is RIP or not. We are interested in designing efficient algorithms whenever possible, and proving hardness results otherwise.

Efficient certification of RIP for *arbitrary* matrices seems hard, being closely related to problems like Small Set Expansion or Densest k-Subgraph. We review a probabilistic construction of RIP matrices, and introduce a planted model for creating adversarial non-RIP matrices, by hiding a sparse set of highly correlated columns. We also consider prior work on the hardness of approximating RIP constants, and present conjectures on the limitations of semidefinite programming.

Kavish Gandhi and Noah Golowich

Analysis of Boolean functions

Mentor: Yufei Zhao

A *Boolean function* is a function of n input bits that outputs a single bit. Such a function can be viewed, from the perspective of social choice theory, as a voting rule for an election with two candidates, in which the n input bits represent the preferences of n voters and the output bit represents the winner of the election. In this context, we introduce several properties of Boolean functions, including influences and noise stability. We then describe how these properties of a given Boolean function, as well as its Fourier expansion, play a role in proving some results, such as Arrow's Theorem and Peres's theorem, that have important implications in social choice theory. Finally, we mention how some of these results on Boolean functions also relate to learning theory and circuit design.

SESSION 2

Dhruv Medarametla

Bounds on the norms of locally random matrices

Mentor: Aaron Potechin

Project suggested by Aaron Potechin

This project deals with exploring and bounding the spectral norms of locally random matrices. Let G be a graph with n vertices labeled from 1 to n chosen randomly from the set of all possible graphs on n vertices labeled from 1 to n, and let H be a bipartite graph with partite sets $U = \{u_1, u_2, \ldots, u_k\}$ and $V = \{v_1, v_2, \ldots, v_l\}$, where $k, l < n$. The locally random matrix of H, denoted by M, is then a $\binom{n}{k} \times \binom{n}{l}$ matrix where each row or column corresponds to a sequence of k or l distinct integers from 1 to n respectively. The entry in row A, which corresponds to the sequence $\{a_1, a_2, \ldots, a_k\}$, and column B, which corresponds to the sequence $\{b_1, b_2, \ldots, b_l\}$, is $(-1)^{E(A,B)}$, where $E(A, B)$ is the number of pairs (i, j) such that edge (a_i, b_j) exists in G and the edge (u_i, v_j) exists in H. We exploit properties of the spectral norm to bound it with the quantity $\text{tr}((MM^T)^k)$, which we then limit by creating an interdependence graph between the individual entries in the product $(MM^T)^k$. This process implies that the spectral norm of M is at most the quantity $n^{\frac{k+l-z}{2}} \log^z n$, where z is the size of the maximum matching in H. In addition, we consider how this bounding allows for the verification of positive-semidefinite matrices, as well as applications in the Sum of Squares Hierarchy.

Karan Sarkar

On modular extensions to Nim

Mentor: Dr. Tanya Khovanova

Project suggested by Dr. Tanya Khovanova

Nim is the basis of combinatorial game theory. It involves a set of heaps of tokens. There are two players who alternate taking away tokens from a set of piles of tokens. A player may take tokens from one pile only. The winner is the player who takes the last token. We investigate a generalization of Nim, which we call an m-Modular-Nim in which a player, in addition to the standard Nim moves, may also take away a multiple

of m total tokens from any subset of piles. We found the winning strategy for this game with 2 piles and a conjecture for the winning strategy for any number of piles and odd m.

Caleb Ji, Robin Park, and Angela Song

Combinatorial games of no strategy

Mentor: Dr. Tanya Khovanova

Project suggested by Prof. James Propp, UMass Lowell

We investigate combinatorial games of no strategy; that is, two-player games where every legal sequence of moves will result in the same winner. To each game we assign a directed graph, with the positions of the game representing vertices and the edges representing moves. By showing the graphs for two different games are isomorphic, we can prove certain combinatorial identities. We also count the number of ways to play games such as Planted Brussel Sprouts, Chocolate Break and Mozes' Game of Numbers, leading to interesting combinatorial results.

SESSION 3

Mehtaab Sawhney

A study of bar and arc k-visibility graphs

Mentor: Jonathan Weed

Project suggested by Jesse Geneson

In circuit design, chips with the minimal number of layers are desired. In order to model this problem Bar Visibility Graphs were created. In Dean et al. Bar Visibility Graphs were generalized to Bar k-Visibility Graphs. Another generalization given by Hutchison of Bar Visibility Graphs was Arc Visibility Graphs and this was extended in Babbit et al. to Arc k-Visibility Graphs. In this talk we discuss an improvement on edge bounds for Arc k-Visibility Graphs and give the first bounds for thickness for Arc k-Visibility Graphs and Semi-Arc k-Visibility Graphs. In particular the Babbit et al. bound of $(k+1)(3n-k-2)$ edges for $n > 4k+4$ and $\binom{n}{2}$ for $n \leq (4k+4)$ in a Arc k-Visibility Graph with n vertices is improved to $(k+1)(3n-k-3)$ edges for $n > 4k+4$ and $\binom{n}{2}$ for $n \leq (4k+4)$. Furthermore upper bounds of $3k+3$ and $2k+2$ are given for the thickness of Arc k-Visibility Graphs and Semi-Arc k-Visibility Graphs. Finally a new problem of calculating the expected number of edges in a random k-Visibility Graph is introduced and the answer is explicitly calculated for Semi-Bar k-Visibility Graphs.

Richard Yi

Continuous model for two-lane traffic flow

Mentor: Prof. Gabriele LaNave, University of Illinois at Urbana-Champaign

Project suggested by Prof. Gabriele LaNave

Modeling the behavior of traffic allows the creation of more accurate simulations, which can lead to improvements in efficiency and safety on the road. In this project, we introduce the variety of models used to describe traffic flow and present a new model from the macroscopic view, drawing upon current ideas in traffic flow theory. We also extend the

Lighthill-Whitham-Richards model to multiple lanes by using the law of conservation of vehicles.

Daniel Guo

An infection spreading model on trees

Mentor: Prof. Partha Dey, University of Illinois at Urbana-Champaign

Project suggested by Prof. Partha Dey

A common study topic nowadays is the study of the spread of infectious diseases and the rate at which these diseases spread. This topic is important because models spread and growth of such diseases can lead to a better understanding and prediction of such instances in real life. As a result, the problem is about the spread of a theoretical infection on different types of trees. The first tree to be investigated was a binary tree. Given a spreading coefficient α, the time that it took to infect to a certain layer of the binary tree was recorded and graphed for different instances. The graph was then analyzed using statistical methods.

SESSION 4

Girishvar Venkat

Signatures of the contravariant form on Specht modules for cyclotomic Hecke algebras

Mentor: Siddharth Venkatesh

Project suggested by Prof. Pavel Etingof

Signatures of representations of the Hecke Algebra of the complex reflection group $G(r, 1, n)$ will be computed. The Hecke Algebra will have the parameter q be a complex number on the unit circle that is not a root of unity. The generic representation theory of the group $G(r, 1, n)$ will be discussed followed by the representation theory of the Hecke Algebra. Then the signatures of the contravariant forms on the Hecke Algebra will be computed.

Samuel Rush

Signatures in representations of rational Cherednik algebras

Mentor: Gus Lonergan

Project suggested by Prof. Pavel Etingof

Here we describe the signatures of a hermitian form created with representations of the rational Cherednik algebras $H_\kappa(D_n)$ of dihedral groups. Particularly, we investigate the values of these signatures in the trivial representation of the algebras associated with odd dihedral groups in which the parameter κ acts as a constant. Notably, we find the particular cases in which the Hermitian pairing of two basis elements of the representation is non-0.

Luke Sciarappa

Algebras in representations of the symmetric group S_t, when t is transcendental

Mentor: Nate Harman

Project suggested by Prof. Pavel Etingof

In this presentation, we explain, motivate and partially answer the question: what are the simple algebras in the Deligne categories $\mathrm{Rep}(S_t)$? The family $\mathrm{Rep}(S_t)$, defined for all $t \in \mathbb{C}$, interpolates the family of categories consisting of the category of representations of the symmetric group S_n for $n \in \mathbb{N}$. We give a classification of simple algebras in $\mathrm{Rep}(S_t)$ for transcendental t.

SESSION 5

Brandon Rafal Epstein

The defect angle and the relation to the Laplacian matrix

Mentor: Prof. Martin Roček, SUNY at Stony Brook

Project suggested by Prof. Martin Roček

Finite triangulations, which are approximations of continuous surfaces that are constructed by tiling them with triangles, have a discrete notion of curvature called the defect angle. The defect angle at a vertex is the value by which the sum of the angles at that vertex falls short of the expected 2π radians about an arbitrary point. We study Γ, a function of the edge-lengths of a triangulation. This function has the property that its variation with respect to an infinitesimal rescaling of all the edges meeting at a vertex by a common factor is proportional to the vertex's defect angle.

We found the second order variation of Γ by deriving the first two terms in its Taylor series, and showed that second-order terms could be written based on a Laplacian matrix. We compared this to other natural constructions of Laplacian matrices on triangulations and showed that the expressions of these operations were equivalent up to proportionality factors. This work has applications in the study of surfaces, especially where a discrete approximation of a continuous surface is required. For example, it might be of use in string theory, whereby the loops of energy that replace the point particles of the standard model trace out surfaces as they travel through space-time. Extensions of this work may show how discrete geometry can be used to encode some aspects of smooth geometry.

Rachel Zhang

Statistics of intersections of curves on surfaces

Mentor: Prof. Moira Chas, SUNY at Stony Brook

Project suggested by Prof. Moira Chas

Every orientable surface with at least one boundary component can be cut into a non-unique planar model, whose edges can then be labeled with letters that read out a surface word. Reduced directed closed curves on a surface are then associated with a curve word which is identified by the letters of the edges through which the curve passes. The length of a curve is defined to be the number of letters in the curve word. Any subword of the curve word corresponds to a segment of the curve.

For constant n and a fixed curve ω on a surface, as n goes to infinity, the distribution of the intersections of ω and all curves of length n appears to approach a Gaussian distribution when normalized. The mean of this distribution grows proportionally with n and can be computed by finding all pairs of subwords of ω and a curve of length n that imply an intersection.

Arthur Azvolinsky

Explicit computations of the frozen boundaries of rhombus tilings

Mentor: Alisa Knizel

Project suggested by Prof. Vadim Gorin

A *rhombus tiling* (aka a *lozenge tiling*) of a polygonal domain is defined as a complete covering of the domain with 60^o-rhombi with neither holes nor overlaps. Computer simulations are used to randomly tile a polygonal domain. As the rhombi get smaller and smaller, ordered facet formations will form along the boundaries. The sharp boundary between these ordered facet formations and the disordered region is defined as the *frozen boundary*. This frozen boundary is always an inscribed curve. The goal of this project, suggested by Professor Gorin, is to explicitly compute the equations of the inscribed curves that are the frozen boundaries.

Using the concept of curve duality, we have developed algorithms for computing the frozen boundaries of two domains: a hexagon with an ellipse for the frozen boundary and an octagon with a cardioid as the frozen boundary. Using these algorithms, we have produced a number of examples of these computations.

SESSION 6

Meena Jagadeesan

The exchange graphs of maximal weakly separated collections

Mentor: Miriam Farber

Project suggested by Prof. Alexander Postnikov

We study the exchange graph $\mathcal{G}^{\mathcal{I}}$ of maximal weakly separated collections inside the positroid $M_{\mathcal{I}}$, where $\mathcal{I} \in \binom{[n]}{k}$ is a connected Grassmann Necklace. It was previously shown by Postnikov, Oh and Speyer that $\mathcal{G}^{\mathcal{I}}$ is connected. We show that if it is a cycle, then it must have length $1, 2, 4$ or 5 and that if it is a tree, then it must be a path. We construct exchange graphs that are paths of all lengths. We also define a notion of C-constant graphs $\mathcal{G}^{\mathcal{I}}(C)$ and establish a connection between these graphs and exchange graphs, showing that the set of C-constant graphs of a fixed co-dimension c is isomorphic to the set of exchange graphs with interior size c. We study the possible orders of $\mathcal{G}^{\mathcal{I}}(C)$ where C is of codimension $c \in \{1, 2, 3, 4\}$ and present a conjecture for the maximal possible order in the general case, relating this order to the Catalan numbers. We prove this conjecture for $c \leq 4$.

Meghal Gupta

Extremal functions of forbidden matrices

Mentor: Jesse Geneson

Project suggested by Jesse Geneson

We will investigate the extremal functions of $0, 1 \ldots k$-matrices, with a specific focus on $0, 1, 2$-matrices. The extremal function of a $0, 1 \ldots k$-matrix P, $ex_k(P, n)$, is the maximum sum of numbers a larger $n \times n$ $0, 1 \ldots k$-matrix can have and still $avoid$ P. We say A $avoids$ P if one cannot take a submatrix of A and then possibly replace entries with smaller numbers to eventually end up with P. In this talk, we will prove some bounds on the values of these extremal functions based on extremal functions of corresponding $0, 1$-matrices, something into which good research has already been done. We also explicitly find the extremal function of some special $0, 1, 2$-matrices.

David Amirault

Better bounds on the rate of non-witnesses of Lucas pseudoprimes

Mentor: David Corwin

Project suggested by Dr. Stefan Wehmeier, MathWorks

Efficient primality testing is fundamental to modern cryptography for the purpose of key generation. Different primality tests may be compared using their runtimes and rates of nonwitnesses. With the Lucas primality test, we analyze the frequency of Lucas pseudoprimes using MATLAB. We prove that a composite integer n can be a strong Lucas pseudoprime to at most $\frac{1}{6}$ of parameters P, Q unless n belongs to a short list of exception cases, thus improving the bound from the previous result of $\frac{4}{15}$. We also explore the properties obeyed by such exceptions and how these cases my be handled by an extended version of the Lucas primality test.

SESSION 7

Jacob Klegar

Tiling-harmonic functions

Mentor: Prof. Sergiy Merenkov, CCNY-CUNY

Project suggested by Prof. Sergiy Merenkov

Tiling-harmonic functions are a class of functions on square tilings that minimize a certain energy. They may provide a useful tool in studying square Sierpiński carpets. We define tiling-harmonic functions and present the main conjectures in this field. Following this is a discussion of our progress towards proving these conjectures, including a maximum modulus principle for tiling-harmonic functions and experimental numerical bounds for Harnack's constant. Finally, we explore comparisons between tiling and graph harmonic functions, especially in regards to oscillating boundary values. Given that these two classes of functions are very similar, and many properties of graph harmonic functions are already known, an understanding of the relationship between the two may prove very useful.

Ahaan Rungta

Mathematically modeling the motion of cells in porous media

Mentor: Andrew Rzeznik

Project suggested by Prof. Jörn Dunkel

Biological cells are known to float, or "swim", in the medium they inhabit. The motion of such swimming cells usually depends on the medium and other external factors such as constant forces created by fields. Many of these systems have been studied in past research. For example, researchers have discovered much about the swimming properties of single-celled organisms in an ideal free fluid with no disturbances. However, many questions are yet to be answered about how cells move in more realistic environments, such as porous media, which is what we study.

To simulate a porous structure for our cell's environment, we write code to simulate random walks for some simple lattice structures within boundaries. Then, to mimic factors that bias the walk, we add persistency to our code. For every scenario, we form frequency graphs which plot the number of times the particle/cell visited every point in our grid. These graphs give us some insights into what might be interesting to study in the future. Ideas on further research include generalizations, even randomizations, of lattices, analogs of diffusion coefficients in random flow, investigation of conjectured phase transitions, and the effects of boundaries on motion microorganisms, both analytically and experimentally.

Nick Diaco

A new coin weighing problem and concealing information

Mentor: Dr. Tanya Khovanova

Project suggested by Dr. Tanya Khovanova

This article of research stems from a question posed by Alexander Shapovalov in 2007 towards a high school audience:

> A judge is presented with 100 coins that look identical, knowing that there are either two or three counterfeits among them. All the real coins have the same weight, and similarly, all the fake coins have the same weight — but are lighter than the real ones.
>
> You yourself know that there are exactly three fake coins and you know which ones they are. Can you use a balance scale to convince the judge that there are exactly three fake coins without revealing any information about any particular coin?

This coin weighing problem is unlike another other traditional balance puzzle in that we already know everything about the coins; our task is to reveal just enough information to show how many of coins are fake without revealing anything else. We will attempt to offer various proofs for this problem along with a series of generalizations, advances, and research objectives.

SESSION 8: MEDICAL INFORMATICS

Ashay Athalye

Machine learning characterization and prediction of intrinsically disordered protein interactions: A focus on BRCA1

Mentor: Dr. Gil Alterovitz

Project suggested by Dr. Gil Alterovitz

The BRCA1 (breast cancer 1 early onset) tumor suppressor gene is involved in diverse biological signaling processes including transcription, cell-cycle checkpoint control, apoptosis, and DNA repair. Mutations in this gene are responsible for approximately 40% of inherited breast cancers and more than 80 % of inherited breast and ovarian cancers. The BRCA1 gene product is an intrinsically disordered protein (IDP) that contains a 1480-residue disordered central region (83% of primary sequence) flanked by a 100-residue structured RING domain at the N-terminal and 250-residue structured tandem BRCT domains at the C-terminal (17%). While more than 400 proteins have been reported in various databases to interact with BRCA1, little is known about interactions to the disordered region: the exact binding-site residues and number of proteins that bind via intermediaries is unknown. Furthermore, the proteins that interact with BRCA1 have not been organized into functional subnetworks, and these subnetworks have not been linked to the various biological activities ascribed to BRCA1.

The first goal of this project is to compile and analyze interactions of BRCA1 and identify its known binding-site residues. This data is crucial to building the prediction algorithm that is based on machine learning. The second goal is to develop algorithms that better characterize and predict IDP interactions (binding and binding-site residues) with increased efficiency and accuracy. These objectives will enable the development of novel drugs that mimic fragments that bind to the BRCA1 or replicate the behavior of the BRCA1 and bind to its partners.

Several databases including Gene Metacore and String-DB were used to compile BRCA1 interaction data, map interactions onto a network, and run enrichment analyses. Python scripts that calculated delta accessible surface area and DSSP (define secondary structure of protein) with a rolling ball method were conducted on BRCA1 complexes in the Protein Data Bank (PDB) in an attempt to infer binding-site residues. The current framework for prediction that implemented supervised multi-task learning with a KNIME workflow of a Weka predictor and Bayesian network was run, outputting a receiver operating characteristic (ROC) curve that described whether or not a partner bound to the molecular recognition fragments (MoRFs) of the protein in question. Results from mapping BRCA1 interactions across several species demonstrated that the BRCA1 protein functions both individually as well as jointly in protein complexes, and that proteins that form functional complexes with BRCA1 also have separate independent functions.

Arul Prasad

The significance of disordered residues in bacterial drug resistance and SNP interactions in relation to disease associations

Mentor: Dr. Gil Alterovitz

Project suggested by Dr. Gil Alterovitz

(1) The rise of multi-drug resistance (MDR) is one of today's biggest issues. Previous research has indicated that intrinsically disordered proteins (IDPs) cause MDR. I focus on the specific amino acid residue range in bacteria for which IDPs' effect on MDR is significant and what this means.

(2) Protein interactions have been studied intensively in the causes of various disease. In these protein interactions there are mutations called single nucleotide polymorphisms (SNPs) which have been found to be disease associated. However, most of these studies have focused on domain-domain interactions. Thus, I look at multiple types of protein interactions in relation to disease, and at for which types of protein interactions SNPs have the most significant effect in causing disease.

Kara Luo

Computer simulation of biosynthetic drug modifications to improve binding activity

Mentor: Dr. Gil Alterovitz

Project suggested by Dr. Gil Alterovitz

Intrinsically disordered proteins (IDPs) are proteins that lack a fixed or ordered three-dimensional structure. These proteins possess a highly flexible, malleable random coil-like structure that allows for adaptation with multiple, distinct partners. When binded with a target, IDPs forego a disorder-to-order transition as they fold in a template-dependent manner. With these characteristics, as well as their common involvement in numerous human diseases, IDPs are considered valuable targets for novel drug designs. However, the current strategies for novel drug development are often costly and ineffective. Here we show that binding activity to the peptide deformylase of *Enterococcus Faecium* is greatly improved through computer simulation of biosynthetic modifications. We found that when tested against an IDP of *Enterococcus Faecium*, the best-performing drug in the database of existing drug molecules has a binding probability of 0.917 and a binding affinity of -7.8. With the biosynthetic modifications, the binding affinity is improved to -10.1, demonstrating strong binding activity with the IDP. Our results demonstrate how computer simulation of biosynthetic modifications to existing drug molecules is able to effectively improve drug activity at IDP binding sites.

Andrew Li

Exploring multi-conformational modeling and flexibility of molecular recognition features in improving drug docking

Mentor: Dr. Gil Alterovitz

Project suggested by Dr. Gil Alterovitz

Pathogens such as the ESKAPE group pose serious threats to the human body if contracted due to their antibiotic-resistant nature. The progression of disease caused by these and other pathogens has been found to be highly linked to intrinsically disordered proteins (IDPs), proteins that lack a rigid tertiary structure. Because of their flexible nature, IDPs are prone to mutation. Additionally, such disordered proteins are easier to bind to since they have a more flexible structure than a normal protein, thus making them relevant for drug targeting. In this study, we model flexibility in binding by generating trajectory distributions of the IDP and docking these conformations with a drug found to be viable. The first major goal of this study was to show that simulation of flexibility prior to binding resulted in a higher and more accurate binding score for the IDP and drug. A matched pairs t-test was performed on the dataset and resulted in strong statistical evidence that the flexibility modeling produces a higher score than do traditional methods. In the process of modeling flexibility, it was also found that the separate analysis of sub-chains of an IDP in drug docking both generally maintained accuracy of docking score and reduced program runtime significantly. A script in Python was written to both dock multiple protein-drug pairs in sequence and to analyze individual sub-chains in this manner. Based on these results, it is clear that flexibility-accounting docking procedures should be developed and formalized for regular use with IDPs.

Daniel Lu

Investigating drug synergy mechanisms of disordered protein-related diseases

Mentor: Dr. Gil Alterovitz

Project suggested by Dr. Gil Alterovitz

This project seeks to achieve drug synergy to target the intrinsically disordered proteins (IDP's) involved with human diseases. In the past decade, IDP's have been brought under ever-increasing scrutiny as they not only represent a new class of proteins previously left unmarked, but they also distinguish themselves for their unique lack of structure and the unusual characteristics this causes. This project continues the focus on the role of IDP's in diseases. Drug synergy constitutes a novel approach to fighting IDP-related diseases as it attempts to capitalize on the unique binding characteristics of IDP's and the idea of synergetic interactions during that binding in designing molecule-level drugs. When focusing on a single protein interaction for a disease, synergy can be achieved either through the use of multiple molecule-like drugs targeting the binding sites of both a disordered disease protein and its partner, or by designing a single drug that could simultaneously target/bind to several different regions. Both methods would have the desired result of incapacitating the disease protein, thus providing a new way of stopping the prognosis of the disease.

Research was first conducted to build up a background in the field, other preparations were completed facilitating the execution of the project, and the given analytic software

was utilized to obtain results. The unique characteristics conferred by disorder in proteins, for instance being more flexible and lacking a definite structure, helped give an understanding necessary to conceptualize the purpose of the project and put each step into its larger context. More specific research was then done to identify which human diseases are primarily associated with IDP's, and also what exactly those IDP's are. After also considering factors such as which diseases offered several distinct IDP's and which ones most commonly affect people, a tentative focus on IDP's related to cancer and neurodegenerative diseases was chosen as the best to start with. While research on the IDPs in these categories of diseases was expanded upon, with programs written to automate processes like attaining files modeling the structure of the protein and finding the interactors of the protein, the software of previous students was optimized and run to obtain results on which drugs are effective candidates.

Laura Braverman and Betsy Pu

Genomic and epigenomic signatures of chromosomal domains

Mentors: Maxim Imakaev, Geoffrey Fudenberg, Nezar Abdennur, Boryana Doyle

Project suggested by Prof. Leonid Mirny

Hi-C is an experimental methodology that quantifies the 3D contacts occurring between regions of the genome inside the nucleus. In 2012, Hi-C data revealed the existence of domains along the genome, with higher contact probability between loci within a domain, and lower contact probability between adjacent domains [1]. More recently, higher resolution Hi-C data revealed pairs of genomic locations with peaks of high contact probability, often located near domain borders. These peaks have been interpreted as the bases of physical loops formed by chromosomes in interphase [2]. In parallel, various functional DNA and epigenomic elements have been mapped onto the 1D sequence of the human genome [3, 4]. While characteristics of domains and loops have been studied, their similarities and differences, and how these may in turn lead to their formation, remain unclear. Our study focuses on similarities and differences between the distribution of functional elements across loop and domain borders.

To investigate the functional differences between chromatin domains and loops, we compared positions of high enrichment of various functional elements and epigenetic states with positions of domain and loop boundaries. We saw strong enrichment of insulators at loop bases, and an even higher fold enrichment of promoters at domain boundaries. We observed no significant enrichment at domain boundaries of architectural proteins CTCF, Smc, and Rad21, previously noted to have high enrichment at loop bases [2]. Additionally, we observed a small valley of depletion in the domain profile for promoters as well as many other states reflecting low architectural protein enrichment within a domain. Our results demonstrate the difference in genomic and epigenomic signatures between loop bases and domains. These results suggest that loop bases and domains are associated with different, but sometimes overlapping, phenomena in the genome.

(1) Dixon, Jesse R., et al. "Topological domains in mammalian genomes identified by analysis of chromatin interactions." *Nature* 485.7398 (2012): 376–380.
(2) Rao, Suhas SP, et al. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159.7 (2014): 1665–1680.
(3) Roadmap Epigenomics Consortium. "Integrative analysis of 111 reference human epigenomes." *Nature* 518.7539 (2015): 317–330.

(4) ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome." *Nature* 489.7414 (2012): 57–74.

SESSION 10: COMPUTER SCIENCE

Amy Chou and Justin Kaashoek

Automating generation of programming problems

Mentor: Rohit Singh

Project suggested by Armando Solar-Lezama

Manually creating many problems that involve specific concepts or skills is often a tedious task for a teacher. We present a system based on the Sketch synthesis tool for automatically generating Python programming problems. We first model targeted programming constructs provided by the user as a subset of Python Abstract Syntax Trees (ASTs) in Sketch using Abstract Data Types (ADTs). We then generate the program along with input/output examples that describe the program. A problem can then be presented to the user as a partially complete program with input/output examples as constraints. We are currently able to generate a simple set of Python problems and are working to extend our grammar to synthesize more complex programs.

Harshal Sheth and Aashish Welling

A garbage collected network stack with CSP threads

Mentor: Cody Cutler

Project suggested by Prof. Frans Kaashoek

Modern day operating system kernels are mostly written in lower level languages such as C. The code of these kernels can be hard to understand, with bugs, memory leaks, memory corruption, and deadlocks. One possible solution to this problem is to write kernels in higher level languages. To demonstrate the benefits of this approach, we implement the network stack kernel subsystem in the Go language. We choose Go because it lends itself to the communicating sequential processes (CSP) programming style. The CSP programming style makes reasoning about the correctness of multithreaded programs easier. We demonstrate that our Go network stack is easier to understand than its counterparts written in C and is also automatically parallelizable.

Gregory Barboy, Albert Gerovitch, and Andrew Gritsevskiy

Mobile health surveillance: The development of software tools for monitoring the spread of disease

Mentor: Dr. Natasha Markuzon, Draper Lab

Project suggested by Dr. Natasha Markuzon

Current monitoring of disease spread by the Center for Disease Control relies on well-verified data. However, there is a delay in data reporting, causing a corresponding delay in epidemic detection. Besides, low geospatial resolution of the data makes highly localized predictions difficult. The goal of this project is to develop software tools to collect localized data and enable monitoring the spread of disease in a small community on a daily basis. We generate several common networks, based on the random,

small world, and scale-free human network models. Then, we use the SIR (Susceptible-Infected-Recovered) model to predict and visualize the spread of disease given a number of parameters, including transmission rate and vaccination patterns. To further enhance the models, we use Census data from FactFinder to simulate epidemics on more realistic social networks, reflecting structures of individual towns or cities. However, Census data does not provide information about individuals' health statuses and daily locations. We are developing a crowdsourcing methodology for collecting users' health status data with precise geolocation accuracy. This online application allows users to update their daily health statuses online, providing a visual map of clusters with increased incidence of disease. In addition, we use analytical methods applied to the collected data for monitoring and predicting disease spread. The developed application will provide better ways for early detection of epidemics, identify places with high concentrations of infected people, and help trace the disease to its origin. We discuss our progress towards these goals, demonstrate our application, explain the concepts we incorporated into our research, and present our plans for the future.

SESSION 11: COMPUTATIONAL BIOLOGY

Michael Colavita

Clustering of pathogenic genes in human coregulatory network

Mentor: Soheil Feizi

Project suggested by Prof. Manolis Kellis

Genetic expression and regulation are critical components of a cell's internal function and responsiveness to external stimuli. Gene regulation is the process by which transcription factors enhance or reduce the expression of other genes. We begin by exploring regulatory networks as a means of representing genetic regulation information in a computationally accessible structure. We then utilize this network to generate a co-regulatory network which captures shared regulatory properties between genes.

We examine the modularity of this newly constructed co-regulatory network by using pathogenicity to classify both transcription factors and target genes. Comparing the resulting modularity coefficients to expected values, we find that several genetic diseases display statistically significant assortative mixing, indicating a tendency for the co-regulation of similarly classified genes and potentially indicating the presence of clustered pathogenic nodes on the co-regulatory network. Furthermore, we discuss initial progress on an algorithm based on spectral clustering to identify clusters of pathogenic genes on this network.

Allison Paul

The inference of directed acyclic graphs using spectral clustering

Mentor: Soheil Feizi

Project suggested by Prof. Manolis Kellis

A gene ontology graph is a directed acyclic graph (DAG) which represents relationships among biological processes. Inferring such a graph using a gene similarity matrix is NP-hard in general. Here, we propose an approximate algorithm to solve this problem efficiently by reducing the dimensionality of the problem using spectral clustering. We show that the original problem can be simplified to the inference problem of overlapping

clusters in a network. We then solve the simplified problem in two steps: first we infer k clusters using a spectral clustering technique. Then, we identify possible overlaps among the inferred clusters by identifying maximal cliques over the cluster similarity graph. Our results show a significant improvement both in terms of the performance and computational complexity compared to the existent methods.

Lalita Devadas

Modelling changes in gene expression using five histone modifications

Mentor: Angela Yen

Project suggested by Angela Yen

Previous work in computational biology has shown that gene expression can be correlated with chromatin features. The Roadmap project allowed us to apply this to a new diversity of data, including samples from primary tissues, which had not previously been extensively tested. We used computational machine learning algorithms to determine the predictive relationship between five histone modifications and gene expression in a variety of cells. Our findings largely agree with previous work which suggested that presence of mark H3K36me3 would be strongly and positively correlated with transcription. As expected, due to the increased complexity and heterogeneity of primary tissue samples, we achieved a lower predictive accuracy but were still able to quantify a predictive relationship. We were able to achieve a high predictive accuracy in cell cultures, providing novel insights into epigenetic regulation of transcription in new contexts.

SESSION 12: COMPUTER SCIENCE

Diana Ding and Cristian Gutu

SecretRoom: An anonymous chat client

Mentor: Albert Kwon

Project suggested by Prof. Srini Devadas

Often times users want to participate in online groups without actually giving away their true identity. There are many existing messaging applications designed to do just this, but they sacrifice true anonymity for speed and performance. A good example is Tor who relies on a series of relays to protect anonymity. Though proven efficient, Tor does not guarantee anonymity in presence of strong adversaries like ISPs and government agencies who can conduct in-depth traffic analysis.

We present a new messaging application that implements renown messaging protocols to guarantee true anonymity while also providing good performance in moderately sized groups. By basing the most computationally intensive portions of the algorithms on one centralized server, our application ensures that clients are not only kept anonymous but also free from any unnecessary computations and communications required by the DC-Nets and Diffie Hellman protocols. Doing so allowed an improvement from the traditional $O(n^2)$ communication efficiency to an $O(n)$ while still keeping all clients fully anonymous.

Akiva Gordon and Krishna Suraj

Improving oblivious RAM protocol through novel eviction and access strategies

Mentor: Ling Ren

Project suggested by Prof. Srini Devadas

We explore and develop new methods for improving the bandwidth and runtimes of Oblivious Random Access Memory Programs. We begin with an explanation of the definition of an ORAM, and address the historical progression of previous ORAMs. We discuss previous ORAMS and the improved efficiency of Ring ORAM over Path ORAM. We also discuss the theoretical implications of Onion ORAM as a new method to surpass the $O(\log N)$ limit, and how it can also be improved with regards to server computation.

We also discuss the current eviction strategies and explain possible improvements upon them including more complex block progression systems to improve tree efficiency and eviction quality. We also investigate possible small compromises of security in favor of larger efficiency and speed rewards.

Henry Liu and Ethan Zou

Time traveling in multicore processors

Mentor: Xiangyao Yu

Project suggested by Prof. Srini Devadas

In a world where data and information are being processed and transferred at an ever-increasing rate, the need for multiple processors has become popular in recent years. With multi-processor machines, computing power is significantly improved, but the problem of sharing memory between cores, or memory coherence, arises. A recently proposed solution that is both simpler and more scalable than the widely used directory coherence, Tardis, uses timestamp counters to logically order memory operations to maintain sequential consistency, as opposed to using physical time. The vanilla Tardis protocol, however, uses long hardware counters for timestamp storage. Additionally, for some applications, Tardis generates a large number of renew requests which consumes the precious network bandwidth. Thus, we propose several optimizations: a timestamp compression scheme to reduce the memory cost of storing timestamps, and several lease predictor protocols to increase efficiency by minimizing the number of renewal requests due to cacheline expiration.

SESSION 13: MATHEMATICS

Uma Roy

Infinity crystals for certain generalized quantum groups

Mentor: Seth Shelley-Abrahamson

Project suggested by Tristan Bozec

To a quiver one can associate an algebraic object known as a quantum group, and much effort has been dedicated to studying the negative portion of the quantum group with the theory of crystal bases. Analysis of the combinatorial structure of crystal bases has provided significant insight into the algebraic structure of quantum groups in classical types. Recently, Tristan Bozec gave a definition of generalized quantum groups that extends the usual definition of quantum groups to quivers with loops at vertices. We attempt to

combinatorially characterize the crystal $\mathcal{B}(\infty)$ associated to these generalized quantum groups in a special case. We provide a conjectural combinatorial characterization of $\mathcal{B}(\infty)$ for a particular quiver through equivalence classes of lattice paths, where representatives of each class are special lattice paths called *steep* lattice paths.

Niket Gowravaram

XYX Algebras

Mentor: Dr. Tanya Khovanova

Project suggested by Prof. Alexander Postnikov

We investigate a type of unital algebras, XYX algebras, generated from graphs. More specifically, we study the dimensions of such algebras and whether such algebras are finite or infinite. We prove that cycle graphs will have an infinite algebra. Furthermore, we also show graphs with a vertex of degree four or two or more vertices of degree three will also have infinite algebra. This severely limits the possible graphs with finite algebras. In addition, we show a bijection between words in the algebra corresponding to the path graph and Dyck paths, thereby proving that the dimension of the algebras corresponding to the path graphs are the well-known Catalan numbers.

Eric Nie

Dual Schubert polynomials

Mentor: Pavel Galashin

Project suggested by Prof. Alexander Postnikov

Postnikov and Stanley introduced dual Schubert polynomials and computed the inverse Schubert-Kotska matrix for 231-avoiding and 312-avoiding permutations. They showed that the coefficients are either $-1, 0, 1$. We show that this is false for σ-avoiding permutations where $\sigma \in \{123; 231; 132\}$ and then make some progress towards proving this for the remaining case where $\sigma = 321$.

SESSION 14: MATHEMATICS

Kenz Kallal, Matt Lipman, and Felix Wang

Equal compositions of rational functions

Mentors: Thao Thi Thu Do and Prof. Michael E. Zieve, University of Michigan

Project suggested by Prof. Michael Zieve

The goal of this project is to make progress on the following two important questions about rational functions:

(1) Which rational functions $a(x), b(x), c(x)$ and $d(x)$ satisfy the equation $a(b(x)) = c(d(x))$?
(2) For which rational functions $a(x)$ and $c(x)$ with rational coefficients does the equation $a(r) = c(s)$ have infinitely many solutions in rational numbers r, s?

We will use tools from various areas of math such as complex analysis, Galois theory and algebraic topology. From these tools, we know that any rational functions solving

either of these two questions would yield collections of positive integers satisfying certain constraints, so we must first find all such collections of positive integers satisfying those constraints. We must then determine which of these collections actually correspond to rational functions, and finally we will determine which of the corresponding rational functions satisfy either of the two original questions.

Arjun Khandelwal and Joshua Xiong

Linear algebra methods in combinatorics

Mentor: Chiheon Kim

We give an exposition of the use of linear algebra to obtain bounds in combinatorial problems, beginning with the simple problems of Oddtown and Eventown. We then show an extension of these techniques to more complicated scenarios using the Nonuniform Modular Ray-Chaudhuri-Wilson Theorem, culminating in a proof of the best known lower bound for the size of an explicitly constructed Ramsey graph.