2012 PRIMES Conference abstracts

Section I: Mathematics

Session 1. Discrete Mathematics I

Christina Chen

Apollonian equilateral triangles

Mentor Nan Li

Project suggested by Prof. Richard Stanley (MIT)

In an Apollonian circle packing, the curvatures a, b, c, d of any four mutually tangent circles satisfy Descartes' equation, $2(a^2+b^2+c^2+d^2) = (a+b+c+d)^2$. For an equilateral triangle and a point P, if a, b, c, d denote the squares of the lengths of the sides of the triangle and the distances from P to the vertices of the triangle, then a, b, c, d, satisfy $3(a^2+b^2+c^2+d^2) = (a+b+c+d)^2$. Define a quadruple of nonnegative integers (a, b, c, d) to be a triangle quadruple if it satisfies this equation. It is easy to verify that if (a, b, c, d) is a triangle quadruple, then $(a, b, c, a+b+c-d)$ is also. This operation and analogous ones for the other elements can be represented by four matrices, and define the triangle group to be the group with these as generators. This work analyzes some properties of triangle quadruples and the triangle group. We show that all triangle quadruples can be reduced to one unique root quadruple, and that all primitive root quadruples are contained in one orbit. We also compute the largest triangle quadruple obtainable at each stage and an approximation for the number of triangle quadruples below a given height. Finally, we prove that the triangle group is a hyperbolic Coxeter group.

Rohil Prasad and Jonathan Tidor

Staged self-assembly

Mentor Jesse Geneson

Project suggested by Jesse Geneson

We examine the staged tile assembly model, introduced by Demaine et al. Using physical constraints on the speed of construction of line segments, we establish lower bounds on their construction and achieve optimal efficiency on their building speed in terms of the number of bins and tiles used. Extending the ideas of this construction provides new methods for building squares. We explore simple constructions in temperature $\tau = 2$ of squares, right isosceles triangles, and general monotone shapes. In addition, we present an optimal construction method for building speed of general radially monotone shapes in temperature $\tau = 1$.

Dai Yang

Halving Lines and Underlying Graphs

Mentor Dr. Tanya Khovanova

Project suggested by Prof. Jacob Fox (MIT)

Consider an even number of points in the real plane, in general position. A line passing through two of these points which splits the remaining points into two sets of equal size is called a halving line. If we treat the points as vertices and the halving lines as edges, we obtain a graph, called the underlying graph. We study various interesting properties of underlying graphs, including operations that preserve the edges, the existence of important graph theoretic structures like cycles and cliques, and the maximum number of edges. The last one has been an unresolved problem in geometric combinatorics for the past few decades, to which we will show an improved upper bound.

Session 2. Discrete Mathematics II

Ravi Jagadeesan and Nihal Gowravaram

Beyond Alternating Permutations: Pattern Avoidance in Young Diagrams and Tableaux

Mentor Dr. Joel Lewis

Project suggested by Dr. Joel Lewis

We investigate pattern avoidance in alternating permutations and generalizations thereof. First, we study pattern avoidance in an alternating analogue of Young diagrams. In particular, we extend Babson-West's notion of shape-Wilf equivalence to apply to alternating permutations and so generalize results of Backelin-West-Xin and Ouchterlony to alternating permutations. Second, we study pattern avoidance in the more general context of permutations with restricted ascents and descents. We consider a question of Lewis regarding permutations that are the reading words of thickened staircase Young tableaux. We give a recursion to enumerate the set of such permutations that avoid 321 and investigate a family of patterns we call "repetitive patterns."

William Kuszmaul and Ziling Zhou

Equivalence classes of permutations generated by replacement sets

Mentors Darij Grinberg and Sergei Bernstein

Project suggested by Prof. Richard Stanley (MIT)

We study a family of equivalence relations in S_n created in a manner similar to that of the Plactic relation studied by Knuth and the Forgotten relation studied by Novelli and Schilling. For our purposes, two permutations are equivalent if one can be reached from the other through a series of pattern-replacements using patterns in an established replacement set. In particular, we are interested in the number of classes created in S_n by each relation and in characterizing these classes. Imposing the condition that the replacement set contain the rotations of a single pattern, we find enumerations for the number of non-trivial classes. When the pattern is an identity permutation, we are able to compare the sizes of these classes and connect parts of the problem to Dyck paths. For the case where the replacement set is of size more than two and consists of patterns of length three, we provide formulas for the number of classes in all previously unsolved cases. Furthermore, we consider redefining the relation to use two replacement sets. We provide enumerations for the number of classes for 9 of the 15 previously unsolved cases and find that several of these enumerations yield the same results as those yielded by the Plactic relation and the Forgotten relation. The reason for this phenomenon is still largely a mystery.

Session 3. Discrete Mathematics III

Aaron Klein

Enumeration of graded poset structures on graphs

Mentor Yan Zhang

Project suggested by Yan Zhang

Studying poset structures on graphs is the well-studied branch of combinatorics known as acyclic orientations. Less well studied, however, are *graded* poset structures on graphs. In this paper, we call a graded poset structure on a graph a *ranking* and study the enumerative aspects of rankings. We use elementary combinatorial methods to count the rankings of some families of graphs and also give a generating function method for the general case (though this method is not necessarily computationally efficient). We emphasize a subset of graphs called *squarely generated graphs* (which include grid graphs), demonstrate an unexpected link between the enumeration of their rankings and their 3-colorings, and use this relationship to generate explicit formulae for certain families of grid graphs.

Alan Zhou

Degrees of regularity of colorings of the integers

Mentor Dr. Tanya Khovanova

Project suggested by Prof. Jacob Fox (MIT)

An equation with integer coefficients is said to be m-regular if whenever all the integers are colored, each with one of m colors, there is a solution to the equation in which all values have the same color. In 1933, Richard Rado found necessary and sufficient conditions for a linear equation to be regular for all positive integers m, and left behind various conjectures on the degrees of regularity for equations. In this project, we develop some results on degrees of regularity for inhomogeneous linear equations. Additionally, we reverse the original premise of the problem and find conditions on linear equations to be regular under periodic colorings.

Session 4. Number Theory

Dhroova Aiylam

Modified Farey sequences

Mentor Dr. Tanya Khovanova

Project suggested by Prof. James Propp (UMass Lowell)

In this talk, we will discuss what Farey Sequences are and illustrate some of their beautiful properties. We then describe a generalized, modified Farey Sequence generated from weighted mediants, and explore some properties of the specific case $k = 3$. We present a set of lemmas and theorems pertaining to this case, along with some conjectures for which there is strong empirical evidence. We will also discuss the fractal-like properties that emerge in the sequence. Finally, we will look at those properties that do and don't generalize to other values of k. We will mention, in particular, the cases where k is prime, and present a theorem and some conjectures specific to these cases.

Saarik Kalia and Michael Zanger-Tishler

Schmidt games and a family of anormal numbers

Mentor Tue Ly

Project suggested by Prof. Dmitry Kleinbock (Brandeis University)

In 1966, Wolfgang Schmidt invented the (α, β) game to prove countable intersection properties of the set of badly approximable numbers. It was shown recently to be a useful tool in dynamical systems and diophantine approximations. However, natural questions about the winning-losing parameters of the some sets have not been studied thoroughly. The purpose of this paper is to partially answer these questions for some interesting subsets of the real numbers.

Session 5. Representation Theory

Sheela Devadas

Modular representations of Cherednik algebras

Mentor Dr. Steven Sam

Project suggested by Prof. Pavel Etingof (MIT)

We study lowest weight irreducible representations of rational Cherednik algebras attached to the complex reflection groups G(m,r,n) in characteristic p. The goal of our work is to calculate Hilbert series and free resolutions of these representations. By studying the kernel of the contravariant bilinear form on Verma modules, we proved formulas for Hilbert series of irreducible modules in a number of cases, and also obtained a lot of computer data which suggests a number of conjectures. We also have conjectured free resolutions for the lowest-weight representations for the groups G(m,m,2) and plan to do so for more groups in the future.

Fengning (David) Ding

Infinitesimal Cherednik algebras

Mentor Sasha Tsymbaliuk

Project suggested by Prof. Pavel Etingof (MIT)

Infinitesimal Cherednik algebras H_α of $\mathfrak{gl}_n$ and $\mathfrak{sp}_{2n}$ were introduced by Etingof, Gan and Ginzburg as continuous analogs of the widely-studied rational Cherednik algebras. For certain deformation parameters α, H_α specializes to different algebras, including $\mathfrak{Usl}_{n+1}$ and $\mathfrak{Usp}_{2n}\#A_n$. Thus, the study of H_α will generalize and connect results from the theories of many algebras. In this presentation, I will describe our progress on the computation of the center of H_α. By considering Poisson infinitesimal Cherednik algebras, we calculated the highest terms of the central elements of H_α. We then found a method to obtain the center of the non-Poisson algebra from the highest terms in certain cases; using this method, we proved a formula for the lowest degree central element of $H_\alpha(\mathfrak{gl}_n)$, which allowed us to classify all finite-dimensional irreducible representations of $H_\alpha(\mathfrak{gl}_n)$.

Section II: Computer Science

Session 6. Algorithms and Complexity

Steven Homberg and Eli Sadovnik

Improving the Efficiency of Fault-Tolerant Distributed Shared-Memory Algorithms

Mentor Dr. Peter M. Musial

Project suggested by Prof. Nancy A. Lynch (MIT)

Shared memory, accessible by read and write operations, provides the usual basis for writing concurrent programs for parallel computers. Shared memory has also been proposed as an abstraction for use in programming distributed networks. In this work we are interested in emulation of shared memory in dynamic distributed networks of heterogeneous, unreliable components, where storage nodes may come and go, nodes may also fail by crashing, processing speeds may vary, and the communication channels may arbitrarily delay messages. In this setting survivability of the shared memory and availability of its content is ensured by the use of replication. However, replication introduces the challenge of maintaining consistency among the replicas while managing dynamic behaviors of the deployment setting. Algorithms that emulate shared memory in dynamic, distributed networks exist, but their implementations are costly. Although the fundamental costs associated with replication cannot all be avoided, an effort should be made to remove any unnecessary redundancy. In this work we focus on an algorithm called RAMBO, introduced by Gilbert, Lynch and Shvartsman. A commonly accepted assumption is that communication is more costly than processing, therefore this work focuses on ensuring that all of the communicated information in RAMBO is fully utilized, thus promoting reduction in the message complexity and the improvement of operation latency.

Ziv Scully

Efficient calculation of determinants of symbolic matrices with many variables

Mentor Dr. Tanya Khovanova

Project suggested by Dr. Ben Hinkle and Dr. Stefan Wehmeier (MathWorks)

Efficient matrix determinant calculations have been studied since the 19th century. Computers expand the range of determinants that are practically calculable to include matrices with symbolic entries. However, the fastest determinant algorithms for numerical matrices are often not the fastest for symbolic matrices with many variables. We compare the performance of two algorithms, *one-step fraction-free Gaussian elimination* and *minor expansion*, on symbolic matrices with many variables. We show that, under a simplified theoretical model, minor expansion is faster in most situations. We then propose new optimizations for minor expansion and demonstrate their effectiveness with empirical data.

Surya Bhupatiraju

On the complexity of the marginal satisfiability problem

Mentor Alex Arkhipov

Project suggested by Prof. Scott Aaronson (MIT)

Consider sets of marginal constraints on an n-dimensional tableau, where these sets constrain the sums of entries in particular 1-dimensional slices. The marginal satisfiability problem (MSP) asks about the satisfiability of these constraints, or whether or not there exists a satisfying assignment of the tableau in which each constraint is met. In probabilistic terms, let desired marginal distributions for every subset S of size c of indices from $[1, \ldots, n]$ be denoted by D_S. MSP asks if there exists a distribution X over n-tuples of values in $[1, \ldots, m]$ such that the S-marginal of X equals D_S. If such a distribution exists, these collection of marginal constraints is said to be satisfiable. The marginal satisfiability problem (MSP) asks whether a polynomial-size collection of constraints given as input, each over at most a constant number k of indices, is satisfiable.

This classical problem is known to be NP-hard by a reduction to 3-coloring, while a quantum variant was shown to be QMA-complete and MSP is already NP-complete for 3-dimensional tableaus, even in restricted sizes of these tableaus. We calculate explicit bounds for computations when various parameters are varied, indicating the uncertainty of whether or not MSP is in NP. We also entirely characterize the 2-dimensional case, providing two constructions for solutions, introduce the notion of local consistency, and partially solve the general constraints and integral problem variants of MSP.

Session 7. Programming Languages and Robotics

Jesse Klimov and Patrick Long

Jeeves

Mentor Jean Yang

Project suggested by Prof. Armando Solar-Lezama (MIT)

Ensuring data security in modern networks is difficult. Jeeves aims to solve this by creating a simpler model for implementing programs which deal with privacy protocols. Jeeves handles sensitive values which can be regulated with policies and level variables. As written in a paper by Yang, Yessenov, and Solar-Lezama, "Jeeves's declarative policies allow the programmer to specify policies at a higher level and allow automatic handling of dependencies between policies." Sensitive variables are not concretized until they are needed. By providing a sensitive value with a context, it can be concretized to produce a definite value. Programmers can specify different policies and levels for data privacy. However, it remains to be determined whether or not this system can scale. We use Jeeves to create a toy social network and to look at the constraints solved when the network runs. Although it is currently in a developmental stage, it is giving us a good idea of the capabilities and pitfalls of Jeeves.

Chris Kaffine

Comparing NARF and SIFT keypoint extraction algorithms

Mentor Jon Brookshire

Project suggested by Prof. Seth Teller (MIT)

An important step in the comparison and analysis of range data is the selection of key points in a point cloud and the characterization of these points in a feature vector for comparison. The goal of this study was to compare NARF and SIFT, two algorithms used to identify key points, to be able to choose the algorithm that found the most stable and effective points. Additionally, both algorithms had parameters whose ideal setting was unclear, so the effectiveness of the algorithms at various parameter settings was also compared. About 150 frames of range data collected along with sensor position information were used for the comparison. The algorithms were run at various parameter settings, and the sensor position information was used to determine if key points in consecutive frames corresponded to each other. This gave information about the usefulness of each key point, which was used as a means of comparison. The results showed SIFT to generally outperform NARF.

Alexander Sekula

Natural language processing for spoken dialog

Mentor William Li

Project suggested by Prof. Nick Roy (MIT)

Spoken dialog systems rely on both speech recognition and natural language processing to function. This report discusses two key components of a dialog system designed to allow residents in a nursing care facility to ask for information and make phone calls using speech. In order to improve the dialog system, we used natural language processing to both improve confidence scoring in the speech recognizer and enlarge its vocabulary to understand more natural input phrases. To improve confidence scoring, natural language features were extracted from a training set of input sentences, and an adaptive boosting algorithm was used to train a final classifier. For the expanding of the vocabulary, online resources (including Wikipedia, Twitter, and the Amazon Mechanical Turk) were used to associate words like "eating" and "morning" with "breakfast." This word association allows the dialog system to correctly categorize more natural sentences, making the dialog system more convenient for The Boston Home's residents. These approaches were found to yield significant improvement over baseline methods: The natural language features were able to classify 67.7% of hypotheses, while the use of online sources for word associations produced substantial decreases in classification error rates on two distinct datasets. This paper is separated into two parts, first the natural language feature extraction, then the word association to expand the realm of the dialog system's understanding.

Session 8. Computational Medicine

Andrew Xia

Integrated gene expression probabilistic models for cancer staging

Mentor Dr. Jeremy Warner

Project suggested by Dr. Gil Alterovitz

The current system for classifying cancer patients' stages has been around for over one hundred years, and with the modern advance in technology, many parts of the system have been outdated. Because the current staging system emphasizes surgical procedures that could be harmful to the patient, there has been a movement to encourage developing a new Taxonomy, using molecular signatures to potentially avoid surgical testing. For my project, my goal is to develop a cancer classifying system using various data to help determine a cancer's stage. So far, I have looked at the cancer patients' T, N, and M stages to determine the clinical cancer stage of the patient, and I plan to work on stage inference through RNA Sequencing data. Through various different methods, I have yielded different accuracies in my data tree function. After successful completion, this method could be groundbreaking in saving costs and increasing efficiency in cancer research and curing.

Skanda Koppula

Prediction-based Bayesian network analysis of gene sets for genome-wide association and expression studies

Mentor Dr. Amin Zollanvari

Project suggested by Dr. Gil Alterovitz

The rapid accumulation of high-throughput genomic and proteomic data has offered unprecedented opportunities for biologists to gain insights on human disease. The project implements a novel prediction-based framework that analyzes SNP data and generates models for disease diagnosis. The framework first identifies patterns of probabilistic dependence between the disease phenotype genes in a set (akin to genes in pathways possibly associated with the disease). The work focuses on the reproducibility (robustness) of results independent of any specific training data sources, addressing the problem by looking at higher level of abstraction - i.e. gene pathways. This allows for accurate analysis of genetic and demographic data. With the proposed method, we first analyzed alcohol dependency - a multi-factorial disease with partial genetic roots - using two sets of clinical data (with 1395 and 1367 patients). The study was able to validate the robustness and predictive accuracy of the framework and generated diagnostic models. Moreover, such models enable us to intervene when the onset of alcoholism appears likely and to better understand alcoholism's underlying biological mechanisms and risk factors. With appropriate clinical data, the developed approach will be used in the future to construct similar predictive models for other diseases.

Peijin Zhang

Identifying Clostridium difficile *in the ICU using Bayesian networks*

Mentor Dr. Jeremy Warner

Project suggested by Dr. Gil Alterovitz

In recent years, the development of large electronic medical record databases has made large scale phenome-based analysis possible. This project explored correlations in the large ICU-based medical database MIMIC II. Through the analysis of patient ICD9 codes in conjunction with records of tested lab values, we were able to develop a phenome map which showed high probability of occurrence of various ICD9 codes in different peak white blood count ranges, a significant one being *Clostridium Difficile* infection. We were then able to further refine the phenotype definition for *Clostridium Difficile* and subsequently extracted patient data along with randomly selected negative controls from the database to develop Bayesian network classifiers using the program WEKA. The resulting classifiers were able to identify both early and hospital acquired *Clostridium Difficile* infection to significantly useful degrees of sensitivity and specificity. In conclusion, our method of visual knowledge representation enables the creation of testable hypotheses with potential implications for predictive medicine. Our use case has demonstrated that through using these methods, it may be possible to predict serious hospital-acquired complications many hours before the onset of symptoms.

Section III. Computational and Physical Biology

Session 9. Computational and Physical Biology

Hao Shen

Star polymers provide insight on Rabl-like chromosome conformations

Mentors Geoffrey Fudenberg and Maxim Imakaev

Project suggested by Prof. Leonid Mirny (MIT)

Yeast chromosomes are arranged in a specific configuration, in which all centromeres are attached to one side of the nucleus. This type of formation, called Rabl formation, can be modeled as a star polymer, which is the formation of several polymers tethered together at a single origin. In this project, we explored equilibrium properties of systems of 2 to 20 polymers of equal length, where the 2 polymer system served as a control. The interaction pattern between chains of a star polymer conformation can be visualized as an average contact map between two of its arms. We obtain contact maps for star polymers with different number of arms, and show that inter-arm contact probabilities are less frequent with an increasing number of arms. We further analyze cross sections of a contact map, and obtain a quantitative description of contact probabilities between monomers that are equidistant from the origin. We also show that interactions between non-equidistant monomers become less frequent with the increasing number of chains. Lastly, we discover that the end of each chain of a star polymer is more mobile that the middle of each chain, and has a higher probability to interact with the origin. We verify our results by comparison of simulated contact maps with theoretical predictions, as well as experimentally measured contact maps of yeast chromosomes.

Ashwin Murali

Global positioning of interphase chromosomes mediated by local chromatin interactions

Mentors Geoffrey Fudenberg and Maxim Imakaev

Project suggested by Prof. Leonid Mirny (MIT)

Despite our current knowledge about the linear genomic sequence, very little is known about three dimensional chromosomal organization within the nucleus. By measuring contacts between chromosomal loci, chromosome conformation capture based techniques provide new insight into three-dimensional genomic organization. It was shown by a genome-wide Hi-C experiment, chromatin can be divided into gene-rich and gene-poor compartments; regions belonging to the same compartment are attracted to each other and tend to colocalize. Another set of microscopy experiments shows that small, gene-rich chromosomes localize in the center of the nucleus, while big, gene-poor chromosomes reside on the nuclear periphery. We intend to bring these two observations together and check where global chromosome positioning can be explained by local chromatin interaction preferences obtained from Hi-C data. Our model intends to determine the organization of these interphase chromosomes by modeling chromosomes as interacting polymers within the nucleus. Previous attempts to reconstruct chromosomal positioning from Hi-C data relied on a set of geometrical constraints, possibly leading to non-physical long-range interactions. We simulate chromosomal dynamics using only short-range chromatin interactions, yielding an unbiased physically-justified estimate of average chromosomal positions.

Boryana Doyle and Carolyn Lu

Local structure of the chromatin fiber arbitrates 3D chromosomal interactions

Mentors Geoffrey Fudenberg and Maxim Imakaev

Project suggested by Prof. Leonid Mirny (MIT)

While promoter regions of DNA are located next to the gene they influence, enhancer regions of DNA can regulate a gene located further away. Consequently, the 3D geometry of DNA must allow the enhancer to come within close proximity of the promoter and gene regions for transcription activity to occur. Previous experiments have shown that these interactions may be disrupted by forming an insulator element between an enhancer and a promoter. An exact mechanism of insulator activity is currently unknown.

We studied models for insulators by performing molecular dynamics simulations. Our first model consists of forming one or two chromosomal loops between an enhancer and a promoter. We showed that in this model the contact between enhancer and gene regions is not blocked; in fact, contact probability increased in most cases. However, the contact probability of the loop with the rest of the genome was decreased, suggesting an alternative model for insulator activity.

In our second model, we explored how local structure of chromatin fiber influences genomic contacts. This model assumes destabilization of a region of chromatin fiber by breaking part of the fiber into an array of connected nucleosomes. We studied how this modification affected contact probabilities of the two surrounding regions. A small decrease in contact probability was observed between monomers in close proximity to the region of loose fiber. These results indicate another possible model for obstructing contact between enhancer and gene regions. This effect may be further explored by varying the properties of the thin fiber to more accurately simulate a flexible chromosomal fiber.