

2010 Claremont Colleges Mathematics REU Operations Research - Statistics Projects



Johanna Hardin, Pomona College
Clustering Microarray Data

Over the past 20 years, computational biology has changed the way we think about biology, genetics, evolution, and medicine. Through advances in technology, we are now able to measure tens of thousands of genes (within mRNA) simultaneously. Other high-throughput (i.e., massive and quick to collect) techniques are also used to measure, for example, proteins or complete genomic sequences.

Such new high-throughput data has provided a wealth of research opportunities for biostatisticians. Standard statistical methodologies break down when applied to microarray data due to lack of distributional knowledge, multiple comparison problems, and lots of irrelevant information.

One particularly important problem in microarray analysis is finding clusters of genes that are co-regulated. That is, we are interested to find genes that are consistently regulated together or regulated in opposite directions (e.g., when one gene is active another gene is always turned off). Most clustering algorithms, however, take an entire dataset and force every observation into exactly one cluster. With microarray data, we expect many of the genes to be noise or irrelevant, so we would prefer those genes to not be clustered.

During the course of the summer, the REU students will investigate various clustering algorithms and their advantages and disadvantages. We will also expand current algorithms to consider only those data values that unambiguously belong to a given cluster.

Interested students should have at least one semester of statistics, preferably a one-year sequence in probability and statistical theory. There is no biology prerequisite for this project; however, interest in biological applications is expected. Additionally, there is no computational prerequisite; however, any programming background is an added bonus and should be addressed in your application.



Susan Martonosi, Harvey Mudd College Disrupting Terrorist Networks

Detecting clandestine members of large covert social networks, such as the al Qaeda terrorist network, is an important counter-terrorism problem. Osama bin Laden and other top leaders of al Qaeda have remained elusive yet charismatic forces in the recruitment of jihadists and the orchestration of attacks on western interests. Here, we propose research that would develop fast heuristics for detecting clandestine members of large covert networks in order to thwart terrorist activity. Such research would also have implications for robust network design, such as for critical infrastructure networks in the U.S.

We represent the communication network of a covert organization by a graph in which the vertices and edges correspond to members of the organization and the communication links between them. We assume that the visibility of a member to intelligence agencies is proportional to the amount of communication that passes through that individual. Thus, clandestine members of the network try not to communicate too much. We seek algorithms for determining vertices to remove from the network to force more communication to flow through the clandestine vertices, making them more visible. We do this by counting the number of communication paths that **must** flow through the clandestine vertex, which we call the load on that vertex. We then identify subsets of vertices whose removal from the graph causes this number to increase.

During this summer REU project, we will combine techniques of social network analysis with network flow and interdiction methodology to (1) identify structural characteristics that make particular classes of graphs amenable to load diversion and (2) use these insights as the basis for fast algorithms or heuristics for selecting vertices to remove in large networks.

Interested students should have some exposure to graph theory or network optimization and comfort with computer programming. A course in operations research and familiarity with Matlab is an added bonus and should be addressed in your application.



Gizem Karaali, Pomona College Game Theory and School Choice

Game Theory is a dynamic and exciting field of mathematics that attempts to describe and model human behavior using mathematical tools. Game Theory has important and interesting applications in many areas including economics, biology, engineering, and computer science.

An emergent subfield of Game theory called Mechanism Design has received attention recently, perhaps because the 2007 Nobel Prize in Economics was awarded for related work. The main idea of Mechanism Design is to consider “metagames” with a game-designer. Games are “designed” with an eye towards creating situations in which the players disclose their private information. This approach allows the modeling of strategic situations in a new way. Novel ideas such as strategy-proofness and fairness come into play as well as standard game theoretic notions like Nash equilibrium and Pareto efficiency.

This REU project will focus on using Mechanism Design to address the School Choice Problem (SCP). The goal in the SCP is to create a matching, or school choice mechanism, (designed by the school district) that allocates available resources (seats in schools) among players (students with parents as agents) subject to district priorities and legal requirements. Our specific goal will be to understand a core set of mechanisms and their properties with a focus on mechanisms currently promoted for School Choice by economists as well as others that might have additional desirable properties. Each student will be expected to pick a mechanism and investigate its strengths and weaknesses. The group will also develop mathematically rigorous definitions of potentially desirable criteria such as fairness and equity and see how the chosen mechanisms score with respect to these notions.

Interested students should have taken a course in linear algebra and at least one mathematics course in which they were required to write rigorous proofs. Previous exposure to game theory or economics is not required. Nonetheless students should address such experience (if they have it) in their applications.



Mark Huber, Claremont McKenna College MCMC for Spatial Data

Spatial point processes model a variety of data, from the locations of an initial outbreak of a disease to where cities appear in a country and the sites of ant nests in an ecology. These are all examples where the data consists of points lying in a fixed region. The points might appear perfectly random to the eye, but deeper analysis can reveal hidden patterns. Typically, this type of data is modeled using what are known as Poisson point processes.

To analyze these point processes, researchers use Markov chain Monte Carlo (MCMC) methods. A Markov chain takes small random steps that build to introduce overall randomness in a system. Everyday examples of Markov chains are shuffling a deck of cards or mixing up a Rubik's cube by randomly turning sides. A Markov chain is rapidly mixing if only a few shuffles are necessary to thoroughly mix things up.

One type of Markov chain for spatial processes is birth-death chains. In these chains points are added to the process, and removed from the process randomly in such a way that the final collection of points mimics the data. In this project we will be looking at an improvement to these chains called birth-death-swap chains, where not only are points added and deleted, they are also moved around. Our research goal will be to show that the new Markov chain mixes faster than the old, enabling analysis of larger data sets in less time. This will be accomplished through a mix of computer experimentation with the algorithm as well as mathematical analysis.

Interested students should have at least one semester of probability, preferably one which has covered the Poisson process. Additional courses such as stochastic processes, or skills such as computer programming ability will be helpful, and should be addressed in your application.