Population Recovery and Partial Identification

Date Tuesday, March 12

Time 3 pm

Location 303 Mudd

Abstract:
We study several natural problems in which an unknown distribution over an unknown population of vectors needs to be recovered from partial or noisy samples. Such problems naturally arise in a variety of contexts in learning, clustering, statistics, data mining and database privacy, where loss and error may be introduced by nature, inaccurate measurements, or on purpose. We give fairly efficient algorithms to recover the data under fairly general assumptions, when loss and noise are close to the information theoretic limit (namely, nearly completely obliterate the original data).

Underlying one of our algorithms is a new combinatorial structure we call a partial identification (PID) graph. While standard IDs are subsets of features (vector coordinates) that uniquely identify an individual in a population, partial IDs allow ambiguity (and "imposters"), and thus can be shorter. PID graphs capture this imposter-structure. PID graphs yield strategies for dimension reductions of recovery problems, and the re-assembly of this local pieces of statistical information to a global one. The combinatorial heart of this work is proving that every set of vectors admits partial IDs with "cheap" PID graphs (and hence efficient recovery). We further show how to find such near-optimal PIDs efficiently.

Time permitting, I will also describe our original motivation for studying these recovery problems above: a new learning model we call "restriction access", introduced in earlier work. This model aims at generalizing prevailing "black-box" access to functions when trying to learn the "device" (e.g. circuit, decision tree, polynomial...) which computes them. We propose a "grey-box" access that allows certain partial views of the device, obtained from random restrictions. Our recovery algorithms above allow positive learning results for the PAC-learning analog of our model, for such devices as decision trees and DNFs, which are currently beyond reach in the standard "black-box" version of PAC-learning.

Based on joint works with Zeev Dvir, Anup Rao and Amir Yehudayoff