Abstract:

Fairness in Automated Classification: A Multi-group Perspective

The increasing reach of algorithmic classification and decision making into our daily lives — examples range from job advertisements to medical decisions to incarceration — has given rise to an explosion of research into the ethics embodied by these algorithms; in a word, are they “fair”? But what is fairness? Can we test for it? Can we achieve it? Are there limits? The burgeoning study of algorithmic fairness attempts to tackle these challenges from a foundational perspective.

One distinction in this literature is between “group” notions of fairness, which ensure that some statistics are equalized between different populations, and the notion of “individual” fairness due to Dwork et al, which requires that “similar” individuals should be treated similarly. Both approaches have advantages and drawbacks: group fairness can be easier to guarantee, but also easier to abuse. Individual fairness provides powerful protections, but requires knowledge of a metric quantifying similarity between individuals, and does not generalize from a sample of training data.

In recent work we suggest a new approach, which we refer to as “multi-group fairness”: providing aggregate guarantees to a large collection of intersecting sets. This approach can be viewed as bridging group and individual fairness notions. This talk will discuss several developments in this line of work, as well as open questions for future work.

Based on joint works with Cynthia Dwork, Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold and Gal Yona.