Proposal:
Fairness-aware machine learning aims to prevent demographic discriminative outcomes of machine learning models. Existing studies focus on debiasing w.r.t. a handful of pre-defined protected groups. However, this can be of concern in the wild for that 1) the bias of a model is hard to be known in advance, and 2) fairness of other demographic groups that are not known to the debiasing-target can be indeliberately harmed. Open target debiasing (OTD) is an approach to tackle this problem, which imposes no specification on the biased groups and the model is expected to achieve fairness over all combined sensitive attributes. To solve the problem of OTD, we firstly plan to extend the concept of demographic parity into OTD to quantify the fairness over all possible sensitive features. An open target fair representation learning framework, named OTIS, is proposed, where provable guarantees via the lens of mutual information are used to achieve parity and maintain the accuracy of downstream classification tasks. Practical instantiations of the objectives are also going to be provided. Finally, we will conduct experiments on two public real-world datasets to demonstrate OTIS to verify the better parity at similar level of classification F1 score compared with the baseline approaches.