Adversarial Signal Processing

Fernando Pérez-González

University of Vigo, Spain

Signal Processing has successfully become an instrumental discipline in addressing the challenges posed by the digital world. An enormous array of applications have emerged that make use of the theory and algorithms developed in several decades of research. However, most of those applications were developed having honest users in mind. Due to rising concerns about security and privacy, and the popularity of cloud-centric services, it has only been recently that different communities in the field have started rethinking designs to account for the presence of a (malicious) adversary. Interestingly, research in these communities has frequently been carried out with little or no mutual interactions. It is not surprising then that similar problems have been solved independently in contiguous areas, with a lack of a unifying view that could benefit from the power of generalization.

The aim of this talk is to present the basic theory of adversarial signal processing, with motivating examples taken from the fields of watermarking, multimedia forensics, traffic analysis, intrusion detection, biometrics, cognitive radio, etc. We will focus on adversarial hypothesis testing, which is arguably the best understood topic. As a fundamental approach we will show how to use game theory to model the available strategies to both defender and adversary. In some cases of interest, it is possible to find an equilibrium of the game which gives the optimum strategies for both parties and the performance that each can achieve. One interesting instance where such equilibrium exists is the so-called source identification game, which has applications in media forensics, biometrics, network traffic analysis and fraud detection, to name a few.

One of the drawbacks of designing an information processing system under the assumption that an adversary is present is that the solutions generally lead to very conservative designs, which in turn yield a bad performance when the adversary is not there. A possibility, to overcome such a problem, is to run a detection meta-test, aiming at detecting the presence of the adversary. A nice example of such an approach, and the resulting race of arms between the meta-detector and the adversary is the case of sensitivity attacks in watermarking, hill climbing attacks in biometrics and the ACRE (adversarial classifier reverse engineering) attack in machine learning, which we will discuss in this talk. We will reveal the striking similarities between these attacks, and how they can be abstracted to build a comprehensive theory. Then we will present some recent results in the design of a meta-detector

aiming at distinguishing malicious queries which are submitted as part of a sensitive attack, from normal queries submitted by non-malevolent users.

Signal Processing in Communications Group, EE Telecomunicacion, Campus Universitario, 36310 Vigo, Spain
fperez@gts.uvigo.es