Title of the thesis

A General Framework of Estimating Information Leakage for Privacy and Forensics Problems with Imperfect Statistical Information

Abstract

In today's digitally-driven world, addressing information leakage stands as a paramount concern, spanning data privacy, corporate security, and national interests. It centers around safeguarding personal data, preventing identity theft, and mitigating financial fraud risks, while also encompassing the protection of corporate secrets and the preservation of national security. The implications are far-reaching, impacting trust, reputation, financial stability, and individual safety.

Information leakage is typically defined as the increased likelihood of an adversary accurately guessing a legitimate user's private data or any related information when they have access to the user's publicly disclosed data. Traditionally, quantifying information leakage has relied on the assumption that the adversary possesses full statistical knowledge of the privacy mechanism. However, this assumption often does not hold in real-world situations, as adversaries frequently lack complete statistical knowledge of the combined statistics of private, utility, and disclosed data. Consequently, the conventional leakage metrics fail to hold operational significance in scenarios where the adversary lacks such information. This underscores the need for leakage metrics that can effectively capture privacy leakage in these real-world conditions.

In this work, we have addressed these limitations and introduced a novel approach to quantify information leakage when the adversary's knowledge of the privacy mechanism is incomplete. We have presented a set of innovative information-theoretic metrics, including average subjective leakage, average confidence boost, and average objective leakage. Additionally, we have analyzed the properties of these metrics and explored the problem of designing the optimal privacy mechanism that minimizes privacy leakage even in the worst-case scenario while ensuring utility of the proposed mechanism. Furthermore, this work extends these metrics to encompass large-scale datasets in a non-Bayesian framework, which is prevalent in contemporary scenarios. We investigate how these proposed metrics can enhance our understanding of adversarial actions and their potential consequences. We rigorously assess the performance of these metrics using real-world numeric and image datasets, employing various machine learning techniques, and provide a comparative analysis with Bayesian inference methods.