Tutorial on
The State of Data Privacy

Dr. Srivatsan Laxman

Microsoft Research India
Email: slaxman@microsoft.com

Abstract

As learning and data mining algorithms mature, we find ourselves increasingly surrounded and reliant on many applications, like search, social-networking or business intelligence. The data in such settings often contain sensitive information of individuals, corporations or governments, and this leads us to the important issue of data privacy. There is a growing concern that algorithms used for analyzing data may also (inadvertently) compromise privacy by revealing specific information about the parties involved. Early work in data privacy established that mere removal or encryption of Personally Identifiable Information in user records is insufficient to guarantee privacy. This led to a sequence of works that tried to formalize a definition for privacy, starting with k-anonymity, and followed by notions like l-diversity, t-closeness and m-invariance. However, all these definitions were broken by a sequence of simple attacks, based either on responses to multiple queries or on suitable auxiliary information available to an adversary. In 2006, Dwork, et al., proposed the idea of Differential Privacy (DP) where, by adding a calibrated amount of noise, it is possible to guarantee that an adversary will learn essentially the same thing about a user, whether or not the user's record was included in the data. The main benefits of DP are that the guarantees are agnostic to auxiliary information and it is possible to precisely quantify the deterioration in DP guarantee under multiple queries (or composition). DP has quickly gained popularity (especially among the theory community) as an important formal notion of privacy with significant potential. Despite its growing success, there are several drawbacks of DP that have prevented its adoption in practice. Foremost among them is that DP adds very high levels of noise to the output, oftentimes leading to unusable query responses. This is because the DP framework assumes the adversary knows almost all the entries of the data base and disregards any possible probabilistic data generation model for the data. This is contrary to what we see in the real world, where data often has strong statistical characterizations and the knowledge of the adversary about specific data entries is often limited. The statistics community has also explored techniques for disclosure control as a privacy-preservation mechanism, but so far, a broad consensus has evaded the privacy community regarding suitability of statistical assumptions under which disclosure control guarantees may be provided. In this tutorial, I will introduce the area of data privacy and highlight the main challenges in this field of research. A wide range of privacy definitions will be covered including k-anonymity (and its variants), Differential Privacy and statistical disclosure control. One of the goals of the tutorial is to bring out the fundamental difficulties in developing formal notions of privacy and the inherent contradictions that exist between privacy and data analysis. A second goal is to analyze the merits and demerits of various privacy definitions that exist today, hopefully throwing light on what needs to be done in-future to achieve formal, yet practical frameworks for privacy preservation.

Bio: Srivatsan Laxman is a Researcher at Microsoft Research India, Bangalore. He obtained his Ph.D. from the Dept. of EE, IISc., Bangalore, in 2006. His research interests are in the areas of pattern recognition and data mining. In particular, his work has focused on various aspects of pattern discovery, efficient algorithms

International Conference on Management of Data
COMAD 2011, Bangalore, India, December 19–21, 2011

for discovering patterns, statistical analysis/significance of patterns in data and the learning/application of generative models based on frequent patterns in data. In the context of data privacy, his research interests revolve around foundational issues in data privacy, privacy definitions, as well as their practical implications.