Chapter 19
Learning With Differential Privacy

Poushali Sengupta
University of Kalyani, India

Sudipta Paul
https://orcid.org/0000-0001-6561-997X
National Institute of Science, Education, and Research, Bhubaneswar, India & Homi Bhaba National Institute, Mumbai, India

Subhankar Mishra
National Institute of Science, Education, and Research, Bhubaneswar, India & Homi Bhaba National Institute, Mumbai, India

ABSTRACT

The leakage of data might have an extreme effect on the personal level if it contains sensitive information. Common prevention methods like encryption-decryption, endpoint protection, intrusion detection systems are prone to leakage. Differential privacy comes to the rescue with a proper promise of protection against leakage, as it uses a randomized response technique at the time of collection of the data which promises strong privacy with better utility. Differential privacy allows one to access the forest of data by describing their pattern of groups without disclosing any individual trees. The current adaption of differential privacy by leading tech companies and academia encourages authors to explore the topic in detail. The different aspects of differential privacy, its application in privacy protection and leakage of information, a comparative discussion on the current research approaches in this field, its utility in the real world as well as the trade-offs will be discussed.

DOI: 10.4018/978-1-7998-5728-0.ch019

INTRODUCTION

Humans gain “knowledge” by inference from raw events, incidents or structured phenomenon. That implies, as long as it is not meaningful or inferred properly, this “raw data” doesn’t become “information” to be inferred that help humans to grab “knowledge” from. Our chapter refers to the definition of knowledge given by (Davenport et al., 1998). It states that from the perspective of an expert with respect to the particular experiences and principles in the right context give a proper structure to asses and integrate new raw data and information. This structure helps to make the transition of the data to “knowledge” - inside a proper intelligent mind. This phenomenon is equally applicable in the daily routine of any organization, processes, norms and practices.

Raw data goes through processes to add contextual meaning in background to become “information”. These processes are heavily prone to defect and danger depending on the nature of the data, its sensitivity and its usefulness towards the organizations. One of the biggest dangers that these “information” and “raw data” can face is “leakage”. Leakage defined as - when an adversary knowingly disclose sensitive information for business purpose to harm an individual, a community or a particular target for his/her own personal satisfaction. The “leakage” can happen in the pre-processing as well as in the post- processing of “information”.

These situations are not desired by any means. Some of the probable solutions are -

- Make sure all endpoints have basic Cyber security systems.
- Use a data backup and recovery solution using encryption-decryption system.
- Clean up the data Storage on the IT Assets after a certain time - window
- Limit user access privileges to only what is absolutely necessary
- Provide Cyber security awareness training to the employees
- *Build security system inherited from the data itself that doesn’t need any kind of third-party affiliation, but robust and fast enough to provide enough privacy and security promise simultaneously.*

All of the promises except the last one needs some audit from a 3rd party who or which can be a potential attacker. Also, these promises need extensive monitoring from a human perspective all the time which is a tough and cumbersome work. The last point in the above solutions is formally known as “differential privacy” that is currently the default trend of privacy solution.

The definition of “differential privacy” (DP) will be discussed in section 4 thoroughly. But as a promise, DP might be thought of as a restriction which filters the leakage of sensitive statistics at the time of the publication of aggregated information, with respect to a database, in the algorithmic level. To elaborate the above promise some examples are discussed below (Wikipedia – Differential Privacy),

Government departments and agencies use DP algorithms at the time of publishing demographic and other types of statistical aggregated analysis report with the assurance of the confidentiality of the survey takers and responses,

Companies use DP algorithms at the time of collecting user behavior, in every step to stop leakage. This measure is also applicable to the internal analysts.

In whatever way the sales numbers of a business are covered in the process of hiding, those numbers might appear when the same process will be done in the total calculation of a vast region that the business belongs to with a combinations of addition and subtraction. DP algorithms diminish those possibilities from the root itself even if the attackers use robust, interactive query system.