Privacy Limit Theorem

Sashank Dara

August 30, 2016

Abstract

With the explosion in the data being collected some believe privacy is dead. With advances in privacy technologies like homomorphic encryption, differential privacy etc. some believe we are closer to the holy grail of privacy applications. *Privacy* and *Utility* are not necessarily to two orthogonal entities. Few argue both cannot exist together and few believe otherwise. The truth is some where in between. This talks tries to debunk many myths surrounding *privacy preserving technologies* like computation over encrypted data, differential privacy etc. Using real world examples, the talk explores the fundamental limitations on how much privacy can be achieved. Maximum utility of the data leads to zero privacy and total privacy erodes the utility of the data completely. The spectrum of opportunity lies in between which modern privacy preserving technologies try to leverage. It is also important to understand the theoretical limitations. This talk explores the fundamental limits of privacy that could be achieved without eroding the utility of the data and conjectures a ***Privacy Limit Theorem***.

1 Privacy

Data is essential for any meaningful insights at the same time privacy is important. Can we have the data and privacy both together ? Yes and No. Lets take a step back to understand what is privacy ?

There are two fundamental aspects to it that are often confused to be same.

1. *Data Privacy:* Hiding the data from the entity that is processing it ?
2. *Identity Privacy:* Hiding the *identity*[1] of an individual in the given data ?

1.1 Data Privacy

Can we encrypt the data and still compute on it without decrypting it ? Computations could be specialised operations like search, sort or advanced operations like machine learning, analytics. Many researchers are building efficient solutions for this problem domain collectively being called as COED (Computation over encrypted Data). It is heavily being sensationalised that its impact on Cloud, IOT and other technologies

[1] Identity could be names, emotions, religious preferences, political opinions etc.

would be humongous. Every advance in areas like *homomorphic encryption, secure multiparty computation, searchable encryption etc* are hailed as a step closer to holy grail. At this moment it is very important to understand the limits of data privacy, collectively this technology could offer (irrespective of the advances on efficiency).

The strongest *data privacy* any traditional encryption scheme could achieve is *adaptive ciphertext indistinguishability (IND-CCA2)* also called as *semantic security*. Informally this means an attacker cannot distinguish pairs of *ciphertext* based on the message they encrypt even if they could interact with encryption/decryption oracles. This is not true for encryption schemes that enable computation over encrypted data. For example, the strongest privacy a *fully homomorphic encryption* scheme can achieve is *non-adaptive ciphertext indistinguishability (IND-CCA1)*. Informally this means an attacker can distinguish pairs of *ciphertext* based on the message they encrypt if they are allowed to interact with encryption/decryption oracles.

So technically encryption schemes that preserve the utility even after encryption offer lesser privacy than schemes that do not preserve any utility. (More realistic examples to be followed in the talk)

1.2 Identity Privacy

Can we hide the identity of the person in a corpus of data ? A promising approach is *differential privacy*. This technology offers a *privacy budget* to hide identity of individuals in *statistical databases* by adding*carefully* defined statistical noise. This helps in querying these databases without providing the ability to link the identity of the person with the query results (thus providing *identity privacy*) within a reasonable privacy budget ϵ.

It is very important to understand the *utility-privacy* trade off here. The maximum *utility* that could be derived from these databases is achieved when no such noise is added. On the other hand random addition of noise would considerably deteriorate the quality of the query results (one could argue that it might offer better privacy since the database is meaningless). Now *differential privacy* offers a well calculated noise levels that retains the utility. The nuance is that not all queries could be performed after adding ϵ (thus constraining the utility).

Lets take an example, a mobile application developer may want to understand the frequency of *emoji*s to improve user experience (like placing them appropriately etc.). The usage could be collected by the app and sent to a remote server for analysis across many users. The precise privacy breach is such data would reveal the individual's emotional state (happy/sad etc.) based on their frequently used *emoji's*. In order to protect the individual's privacy, *differential privacy* could be used and certain amount of calculated noise could be added to the frequency of the usage.

At very high level, if the noise added is too high the privacy of individual could be protected well but the frequency of usage gets distorted and cannot be used. If the noise added is too low the mobile application provider could breach the privacy of the user. Irrespective of the level of noise, it is very easy to determine the emotional state if at all the data is of any real use.

Figure 1: Conflict and Window of Opportunity

2 Privacy Limit Theorem

Privacy and *Utility* are considered to be two orthogonal entities. Either we could achieve complete privacy of the data by encrypting or we have complete utility of the data. The talk conjectures the modern privacy preserving technologies have a only a small window of opportunity to achieve real world privacy.

3 Conclusions

The privacy utility trade off is conjectured as *privacy limit theorem*. Informally it states that the utility erodes while increasing levels of privacy is achieved and privacy erodes when the utility that could be derived from data increases. The spectrum of opportunity for privacy preserving technologies is some where in between.