



Succinct Big Data Representations for Privacy and Efficiency

Researchers

Prof. Shlomi Dolev
shlomidolev@gmail.com

Prof. Ehud Gudes
ehud@cs.bgu.ac.il

Philip Derboko
philip.derbeko@gmail.com

Prof. Jeffrey Ullman
ULLMAN@CS.stanford.edu

Development Stage and Status

Current research is focused on building a model for statistical queries with wavelets and polynomials that keep the data in privacy-preserving way.

Further research includes prediction of the data trends, based on the periodic behavior of the input and extrapolating the data at Fourier's domain, where the Byzantine data [e.g., anomalies or out-layers] is identified and discarded.

Goals

Consider the task of representing information in a privacy preserving and an error-tolerant way by a succinct model, such that it can be formulated even if it contains noise or even if the data are partially corrupted and destroyed. The research group is presenting the concept of data interpolation in data aggregation and representation, as well as in the new big data challenge, where abstraction of the data is essential in order to understand the semantics and usefulness of the data.

The researchers are developing a means of creating a succinct, similitude representation of the data, such that it allows for statistical calculations and also preserves privacy of the data. This dramatically reduces the effects of data breaches and also makes it easier to keep the data safe. The data cannot be breached if it is not there to begin with! These methods are especially valuable when applied to IOT devices, where it saves power, protects storage and networks and keeps the data more secure.

Description

The proposed methods vary with the type and the target queries, and include capturing the essence of the data by abstract function, such as wavelets and polynomials and using controlled sampling to build a compact representation of the data with bounded error. The researchers process the data points to build a model for interpolation, extrapolation and dynamic representation of the data. Those objectives are challenging, since in practice the data can be noisy and even Byzantine, where the Byzantine data represents an adversarial value that is not limited to being close to the correct measured data.