Data Processing under Security and Privacy

Vladik Kreinovich

Department of Computater Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
vladik@utep.edu

1. Need for Processing Large Amounts of Data

- Modern computers allow us to store and process large amounts of the data.
- There is hope that, e.g., by analyzing these huge amounts of medical data, we will be able:
  - to better understand causes of different diseases,
  - to better trace which cure is better for each patient,
  - and as a result, to arrive at more personalized medical practices.
- Current data analytics has indeed led to several interesting medical breakthroughs.
- Similar interesting results have been obtained in agriculture, in biology, in social sciences, etc.

2. Potential Problem with Privacy

- We do not know a priori what kind of dependence we are looking for.
- So, it make sense to allow researchers to ask all kinds of statistical questions.
- However, even with anonymization, this may lead to a violation of privacy.
- For example, we may want to analyze how blood pressure depends on the closeness to I-10.
- We compute the average blood pressure in all the houses on Univ. ave. up to 501, and then up to 503.
- We will thus be able to reconstruct the blood pressure of a person living in 503 Robinson.

3. How Can We Preserve Privacy and Still Allow Researchers to Analyze Data

- If we allow researchers to ask all possible questions based on the exact data, then privacy is not preserved.
- We do not want to limit possible queries: this may prevent us from finding the actual dependence.
- So we should modify either directly the query result or the data values on which this result is based.
- For example, we add random noise either directly to the query result or, indirectly, to the data values.
- Now we have a new problem: the added noise affects the results of data processing.
- We therefore need to know how big is this effect, i.e., how accurate are the results.

4. Probabilistic Methods May Lead to Loss of Privacy

- If we add random noise to the query, then:
  - by asking the same query many times and averaging the result,
  - an adversary will be able to get the exact value
  - and thus, to gain supposedly protected information.
- If we add noise to the original data, then the adversary can take into account that:
  - the same person is usually listed in many different databases,
  - these values are obtained by adding different instances of random noise,
  - so, we can average and reconstruct the actual data.

5. Storing Ranges Instead of Actual Values

- These problems can be avoided if, instead of the actual value, we only store a range of values.
- Example: instead of the exact age, we only mark whether it is:
  - between 20 and 30,
  - between 30 and 40, etc.
- In other words, we only keep an interval such as $[20, 30]$.
- This is the only data stored in the database, so this only the interval can be reconstructed.
- Instead of the original algorithms, we thus need new algorithms – for processing intervals.
- Designing such algorithms is the main focus of our privacy-related research.