CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 32 – November 15, 2023

Prof. Samson Zhou *Scribe: Shima Salehi*

Privacy and Confidentiality

In various scenarios, safeguarding sensitive information is crucial. When releasing databases, ensuring that sensitive information remains protected is a primary concern.

Private Data Analysis

Analysis of diverse datasets, such as medical data for predicting potential issues, pattern detection within social networks or epidemic spread, and utilization of US Census information for apportionment, requires a careful balance between extracting necessary information and preserving privacy.

Consider having two datasets: a Sensitive dataset and an Anonymized dataset. The challenge lies in presenting these datasets while preventing the release of sensitive information. For instance, suppose one dataset includes features like age, zip code, employer, and a "dog" indicator, while another dataset contains name, age, gender, and employer information. Merging these datasets might reveal the names of individuals who own a pet, potentially violating their privacy.

Moreover, privacy concerns emerge in cases like Netflix predicting user preferences for movies using causal inference. Another example involves Netflix canceling certain content due to privacy concerns. Anonymized Netflix data and incomplete IMDb data, when combined, can inadvertently lead to the identification of Netflix data, highlighting the intricacies of data privacy.

It is often believed that releasing more information results in greater accuracy. However, this practice poses challenges to privacy.

Ad-hoc privacy measures like anonymization or deidentification often fall short. Publishing numerous queries with high accuracy on sensitive databases may compromise their privacy. Hence, there's a necessity for a formal mathematical notion to measure privacy.

Possible Notions for Privacy

Notion 1: "The data analyst cannot learn anything about Alice"

Consider a scenario where Alice is known to be an "aggie." An analyst assumes that most Aggies like Reveille. Does this violate Alice's privacy? Not in this case, as "aggie" isn't sensitive information.

However, if a survey concludes that most Aggies like dogs and Alice is a known Aggie, leading an analyst to infer that Alice is more likely to be a dog owner, privacy concerns arise. Even if Alice didn't attend the survey, her Aggie status becomes public knowledge.

Notion 2: "A study is private if the data analyst gains almost no additional information about Alice from the study compared to conducting the same study without Alice's data"

Stability is crucial: a data analyst should reach similar conclusions if any individual data point is replaced by another from the population.

Differential Privacy

Differential Privacy, introduced by DMNS06, establishes a standard for safeguarding privacy in data analysis. An algorithm $A : \mathcal{U}^* \to \mathcal{Y}$ is considered (ε, δ)-differentially private if, for every neighboring frequency vectors f and f' and for all subsets $E \subseteq \mathcal{Y}$, the following inequality holds:

$$\mathbb{P}[A(f) \in E] \leq e^{\varepsilon} \cdot \mathbb{P}[A(f') \in E] + \delta$$

Here, ε (epsilon) and δ (delta) are parameters. For small ε values, e^{ε} approximates $1 + \varepsilon$. This definition aims to maintain the probability of an event E under algorithm A with minor differences in input datasets f and f', ensuring privacy by controlling the impact of individual data points on the output.

The value of δ represents the probability of the mechanism "failing" to achieve differential privacy. If $\delta = 0$, the mechanism satisfies pure differential privacy. For $\delta > 0$, it satisfies approximate differential privacy.

Achieving differential privacy through deterministic algorithms is challenging unless they function as constant functions, indicating that randomness is pivotal in preserving privacy.

This criterion offers a formal framework for assessing and ensuring privacy in data analysis, emphasizing the balance between information utility and protecting individual privacy.