Differential Privacy

Differential privacy is a mathematical method for quantifying and controlling the risk of identifying individual data in a dataset. It provides a way to share statistical information about a group of people without compromising the privacy of any individual. This is achieved by adding a small amount of noise to the data, which makes it impossible to tell whether a particular individual's data was used in the calculation.

The technique is inspired by the fact that data that has been locked up in a mathematical safe cannot be used for scientific research, aggregated for statistical analysis, or processed to train machine learning algorithms.

A strong differential privacy algorithm can enable all these things and more. It simplifies and secures sharing (at least until suitable, efficient homomorphic methods arise).

Why doesn't Anonymization suffice?

The Netflix Prize was an open competition for the best collaborative filtering algorithm for movie recommendations. The dataset released was anonymized, without the users or the films being identified except by numbers assigned for the contest. Such anonymized movie records were published by Netflix as training data for the competition.

However, there were several users who could be identified by linkage with the Internet Movie Database (IMDb) which was non-anonymized and publicly available. Researchers Arvind Narayanan and Vitaly Shmatikov, at the University of Texas at Austin present their studies in their work Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset).

Therefore, such linkage attacks can be used to match “anonymized” records with non-anonymized records in a different dataset.

Differential privacy aims at neutralizing such linkage attacks. As Differential Privacy is a property of the data access mechanism and is unrelated to the presence or absence of auxiliary information available to the adversary.

Therefore, access to the IMDb would no longer permit a linkage attack to someone whose history is in the Netflix training set than to someone not in the training set.

How does Differential Privacy works?

Differential privacy involves adding a privacy loss or privacy budget parameter, typically represented as epsilon (ε), to the dataset. This parameter controls the amount of noise or randomness that is introduced to the raw dataset.

Let's say you have a column in your dataset with "Yes"/"No" responses from individuals. For each individual, you flip a coin. If it lands on heads, you leave the response as is. If it lands on tails, you flip the coin again and record the response as "Yes" if it lands on heads and "No" if it lands on tails, regardless of the actual response.

This process adds randomness to the data. For large datasets and with knowledge of the noise-adding mechanism, the dataset remains accurate in terms of aggregate measurements. However, each individual in the dataset can plausibly deny their actual response due to the randomization.

Noise-adding algorithms in real-world applications are more complex than coin flipping and are based on the epsilon parameter, which balances privacy and data utility. A higher value of epsilon means more accurate but less private data.

Differential privacy can be implemented locally or globally, with noise added to individual data before it is centralized in a database or to raw data after it is collected from many individuals, respectively.

Why is it important now?

Differential privacy is important for businesses because:

It enables businesses to adhere to data privacy regulations like GDPR and CCPA while still being able to analyse customer behaviour effectively. Non-compliance with these regulations can lead to severe penalties. As per a recent report by DLA Piper, a renowned international law firm, fines totalling €273 million have been imposed under GDPR since May 2018. Considering the scale of GDPR compliance and the size of the European economy, these fines are relatively small. However, they are anticipated to rise as countries adopt more comprehensive and automated methods to assess GDPR compliance.

Differential privacy is used in a variety of applications, including:

Census data collection: Governments can use differential privacy to collect census data without revealing sensitive information about individuals.

Medical research: Researchers can use differential privacy to share data about patients without revealing their identities.

Targeted advertising: Companies can use differential privacy to target ads to users without revealing their personal information.

Differential privacy is a powerful tool for protecting privacy in the age of big data. It is becoming increasingly important as we collect more and more data about individuals.

Here are some of the key properties of differential privacy:

Bounded risk: The risk of identifying an individual is bounded by a parameter called the privacy budget. The privacy budget can be adjusted to trade off privacy for accuracy.

Compositionality: The privacy guarantees of differentially private algorithms can be composed, which means that a sequence of differentially private operations will also be differentially private. This makes it possible to combine differentially private algorithms to perform more complex computations.

Resistance to attacks: Differential privacy is robust to a variety of attacks, including attempts to re-identify individuals by linking data from multiple sources.

Differential privacy is a rapidly evolving field, and there is a lot of active research on developing new differentially private algorithms and improving the privacy guarantees of existing algorithms.

Pros and cons

Collaboration and sharing

Pro: Collaboration is necessary. Collaboration is becoming increasingly important in initiatives. Cloud computing is becoming increasingly popular. Finding strong privacy-protection algorithms allows more users and partners to collaborate with data without exposing personal information. Adding a layer of noise also increases safety.

Con: Is it a good idea to share faulty data? It's wonderful to exchange data, but can providing incorrect information help? Differential privacy algorithms function by adding noise, which is another way of stating "error." For some algorithms, such as determining the mean, mistakes can cancel one other out while still producing correct results. More complicated algorithms are less fortunate. Furthermore, when the data sets are tiny, the impacts of the fuzzing might be considerably more severe, resulting in the possibility of large distortions.

Controlling trade-offs between privacy and accuracy

Pro: Good algorithms manage trade-offs. Differential privacy algorithms do more than just add noise. They depict and formalize the choices between precision and privacy. They provide us a knob to modify the fuzzing to our liking. The algorithms allow us to define a privacy budget and then spend it as needed throughout the data processing phases. If you recall mathematics, the approach is attempting to simulate differentiation and determine the slope of the privacy loss.

Many differential privacy algorithms refer to this privacy parameter by the Greek letter epsilon and use it in an inverted manner, with big values of epsilon causing essentially little change in the data and tiny values of epsilon causing enormous quantities of noise to be added. Because of the inverse connection, the number may appear ambiguous.

Con: The fact that Epsilon is still merely a number is a disadvantage. All the mathematical jargon and sophisticated formulae, however, serve only to obscure the fact that a number must be chosen. Is two better than one? Which number is correct? How much is too much? Why not try 1.4232? There is no simple roadmap, and best practices have yet to emerge. Even if they do, can you be certain that the optimal number for, say, the hamburger stands down the street is the correct epsilon value for your garden tool company?

New methods of data protection

Pros: The philosophical approach of differential privacy is a plus. It's not a specific algorithm. There are dozens of algorithms, and researchers are constantly fine-tuning new ones. Some match the strict mathematical criteria, while others come close and provide what some academics refer to as "almost differential privacy." Because each method provides somewhat different assurances, there are several options for securing your data to consider.

Cons: There are no assurances. The differential privacy vision does not provide hard guarantees, just statistical ones that the difference between actual and fuzzy data is confined by some epsilon-governed threshold. So, some true information will leak out, and the noisy version is frequently near, but at least we have some mathematical bounds on how much information is leaking.

Acknowledgement:

www.youtube.com. (n.d.). *Differential Privacy: What? So What? Now What?* [online] Available at: https://www.youtube.com/watch?v=NRf6sUk1bv0

says, E.J. (2021). *Differential Privacy: What it is, how it works, benefits & use cases*. [online] research.aimultiple.com. Available at: <https://research.aimultiple.com/differential-privacy/>.

CSO Online. (n.d.). *Differential privacy: Pros and cons of enterprise use cases*. [online] Available at: https://www.csoonline.com/article/570203/differential-privacy-pros-and-cons-of-enterprise-use-cases.html.

‌

www.youtube.com. (n.d.). *Deep Learning with Differential Privacy (DP-SGD explained)*. [online] Available at: https://www.youtube.com/watch?v=oNSelFJnPaM&t=51s

‌

www.youtube.com. (n.d.). *The Mathematics Behind Differential Privacy*. [online] Available at: https://www.youtube.com/watch?v=QJ3D4koSc6A [Accessed 17 Dec. 2023].

‌

‌

‌