

Analysis of Privacy Leakage Risks in the Context of Security Threats Associated with Digital Device Fingerprinting

Ivan Podsevalov(✉), Alexei Podsevalov, and Vladimir Korkhov

Saint-Petersburg State University, St. Petersburg, Russia
i.podsevalov@spbu.ru

Abstract. In the modern field of data processing, operations are often conducted within distributed systems, where information of individual users is shared across multiple servers. In such a context, users aim to protect their privacy despite the inevitability of transmitting their device parameters to servers, which occurs almost automatically. One of the main risks is the transmission of device hash identifiers (so-called "digital fingerprints") along with private statistical information, which can lead to the near-complete de-identification of users by external organisations. This risk is exacerbated by the potential use of additional statistical data that can be associated with a user.

This study proposes an analysis of existing methods for assessing the level of privacy and the effectiveness of various algorithms in maintaining differential privacy. The paper explores approaches to assessing the leakage of private data in scenarios where the security threat is associated with the computation and transmission of digital fingerprints of devices, representing a significant interest in developing more reliable data protection methods in distributed processing systems.

Keywords: Differential privacy · Security · Privacy leak · Statistics · Fingerprint

1 Introduction

In the big data era, advancements in information technology have significantly improved the quality of services and living standards, thanks to the growth in available information volume and its processing. However, with the increased accessibility of data, there arises a heightened necessity for the protection of personal information privacy. This challenge compels governments and organizations worldwide to develop and implement effective data protection mechanisms, underscoring the crucial significance of privacy in today's information society.

Many intergovernmental organizations are actively working on preserving personal privacy and have established operational regulatory and legislative acts.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14821, pp. 386–404, 2024.
https://doi.org/10.1007/978-3-031-65308-7_27

The European Union is a pioneer in the realm of individual data privacy, having adopted the General Data Protection Regulation (GDPR) in 2016 [1]. Following this, several countries outside of Europe have progressively introduced and enforced privacy laws or regulations, such as Brazil's General Data Protection Law (LGPD) [2], India's Personal Data Protection Bill (PDPB) [3], and Russia's "Personal Data Protection Law".

Despite these efforts, administrative methods of protection sometimes fail to produce the desired outcome. For example, in 2018, the British analytics company Cambridge Analytica accessed Facebook (Meta) user data, resulting in the leakage of at least 80 million units of user information [8]. The data of approximately 500 million customers of the world's largest hotel chain, Marriott, were stolen from its reservation database [9]. Security Magazine published a list of the "best" Top-12 data breaches in 2019 [10], which included the largest American real estate title insurance company, First American Financial Corp, that exposed transaction data of about 885 million individuals, and job seekers in China.

A data breach by the developer and manufacturer MongoDB led to the disclosure of more than 200 million job candidate records. Such information leakage incidents have been occurring one after another in recent years, causing people to grow increasingly concerned about the security of their personal information. These data losses are related to human factors; however, these leaks become a rich source of auxiliary information.

It is evident that relying solely on laws and regulatory acts is insufficient for maintaining confidentiality; hardware and software support for privacy protection methods is required. Countless methods for assessing privacy preservation have been developed, based on maintaining anonymity (for example, k-anonymity [5], l-diversity [6], and t-closeness [7]), which require initial assumptions about specific attacks and knowledge of the volumes of protected data. The research covers ever-new aspects of privacy preservation, considering various combinations of threats.

Academic research in the field of personal data privacy preservation continues unabated as new threats to confidentiality and issues in data science applications emerge. It can be asserted that one of the most successful concepts for assessing confidentiality is the concept of differential privacy [11,15,16]. Introduced by Cynthia Dwork, Frank McSherry, and others [15], this concept ensures that an individual participating in a process that involves further data processing will not face significantly increased risks due to their participation. This guarantee is achieved by declaring that the result of data processing does not substantially (in a controlled manner) depend on whether each individual participates or not.

Thus, differential privacy is a mathematically rigorous definition of privacy compliance conditions, suitable for analyzing large data sets and equipped with a formal measure of privacy loss [11,15]. Differential privacy algorithms strictly take an ϵ parameter as input, which limits the permissible loss of confidentiality for any algorithm execution and offers a specific trade-off between privacy and utility. The concept of cumulative privacy loss has been proven correct, as well

as the possibility of post-processing under the same conditions. Previously, we discussed the theoretical basis for spending the differential privacy budget and proposed an architectural solution [37]. Currently, an extensive theoretical basis has been developed for a large number of relaxed requirement protocols, the so-called (ϵ, δ) - differentially private protocols, guaranteeing a privacy loss of no more than ϵ, with a probability not exceeding $1 - \delta$. Typically, δ is considered "cryptographically" small, this relaxation often provides asymptotically better accuracy than pure differential privacy at the same ϵ value, even when δ is very small.

On the other hand, differential privacy, as originally conceived, does not preclude the possibility of privacy breaches given associative data. In other words, an adversary could exploit correlations between data fragments to compute confidential information about an individual using special mechanisms, as justified in the work "On the Free Lunch" [23] and well demonstrated in "Dependence Makes You Vulnerable: Differential Privacy Under Dependent Tuples" [24]. This work describes how a curious person, by gaining access to a record containing a list of social network friends, virtually obtains a complete probabilistic assessment of user connections.

In their paper [26], Cynthia Dwork and co-authors proved that under certain conditions, no DP mechanism that offers some non-zero utility can prevent the disclosure of information if adversaries possess auxiliary information about the secret data of the storage. This is because an adversary can use auxiliary information to disclose more information than the privacy mechanism assumes. An illustrative example is given in [26]: suppose the exact height of every individual is a secret, and consider a database containing measurements of people's heights across different nationalities. Suppose differentially private statistics are provided - average heights of women of various nationalities. If an adversary has auxiliary information that a certain individual is two inches shorter than the average Lithuanian woman, they can determine that person's exact height by simply subtracting two numbers. This relatively simple example turns into a significant problem under certain conditions.

One such significant condition is the prevalence of various digital devices that can be confidently identified. Fingerprinting or digital fingerprint is a device identification method that uses unique characteristics and parameters of the device, such as configuration, browser data, or hardware features, to determine and track the device. This method can be used for various purposes, including analytics, security, and advertising, but it can also raise questions about user privacy and confidentiality, as it may allow identifying them without their explicit consent.

Moreover, obtaining a digital fingerprint from a device can be accomplished by various methods and in several forms. Here are a few common ways: browser fingerprinting, device fingerprint, IP address and network characteristics, cookies, mobile device fingerprint, and app fingerprint. This topic is well-explored in many works, for instance, in [27,28], and is vigorously used for different purposes.

2 Problem Statement

The study begins with an analysis of the contemporary digital environment characterized by the widespread presence of devices and software capable of tracking users' actions (e.g., sensors in smartphones and trackers in browsers). These tools enable constant monitoring of users even without the transfer of their data. Such observation is stochastic, allowing it to be assessed and analyzed from a scientific standpoint.

This research aims to systematize and study methods for evaluating necessary data perturbations before their transmission to ensure confidentiality within the Local Differential Privacy (LDP) framework. Special attention is given to analyzing entropy, mutual information, and the data transmission channel capacity in scenarios where interaction between the client and server occurs under uncertainty and the potential presence of additional information from potential adversaries.

The main research directions include:

1. Investigating entropy assessment algorithms in various notations, considering privacy and data confidentiality requirements not only in untrusted dialogues but also in scenarios with a known scale of privacy leakage;
2. Examining existing assessments of potential privacy loss for aggregated representations of digital fingerprints of devices or browsers in both one-try processes and interactive dialogues;
3. Exploring various methods of applying LDP to assess privacy loss and determine their effectiveness in the context of this problem;
4. Studying the impact of noise addition algorithm parameters, and consequently, changes in privacy levels on the accuracy of privacy loss estimation;
5. Identifying metrics and methods for evaluating the quality of privacy level assessments in untrusted dialogues.

The central goal of this work is to research and test reliable methodologies for assessing the risks of private information leakage. These methodologies aim to ensure a high level of data protection. Special focus is placed on conditions of untrusted data exchange between the client and server and the use of digital fingerprints of various origins.

3 Problem Relevance

The issue addressed in this research is highly relevant in the contemporary data processing and privacy sphere. Here are several key aspects that underscore the relevance of this problem:

– **Data Privacy Protection**. In the context of collecting anonymous statistical data, protecting users' privacy is critically important. Adhering to data privacy is both a moral and legislative duty, becoming increasingly significant in today's information society.

- **Untrusted Scenarios**. In the client-server model's untrusted dialogue conditions, where neither party fully trusts the other, there is a need for mechanisms to ensure data privacy. This need may arise, for example, from using cloud services or other third-party platforms where users do not have complete control over their data processing.
- **Utilization of Fingerprinting**. A device's digital fingerprint represents unique identification information that can be used to link anonymous data across different sessions or steps in a dialogue. Relying on digital fingerprints can yield valuable statistical insights, but it is essential to ensure the anonymity and privacy of the fingerprints' owners.
- **Local Differential Privacy**. The LDP methodology allows for the collection of aggregated statistical data with minimal disclosure of sensitive information. This problem explores the application of this methodology for data entropy assessment, opening new possibilities for privacy preservation in untrusted scenarios.

Collectively, the issue is relevant due to the need to collect anonymous statistical data, link data using digital fingerprints without disclosing the owner's information, and assess entropy in untrusted client-server dialogue scenarios within the LDP framework. Solving this problem will achieve a balance between utilizing valuable information and protecting user privacy, which is crucial in modern data processing.

4 Device Digital Fingerprint

Websites often need to track visitors for various reasons, such as maintaining user sessions or implementing shopping cart functionality. The traditional method of tracking a user's identity is based on cookies, but in recent years, cookie technology has been revealing more and more problems. For instance, many websites misuse cookies for advertising purposes. These issues have led to growing mistrust towards cookies, prompting many people to install cookie-cleaning extensions or simply use private browsing modes. Concurrently, the technology of capturing digital fingerprints of devices or browsers has been gradually developing and is becoming the new primary technology for tracking users. Let's focus mainly on working with browsers. A browser fingerprint is a collection of all the feature information that can be gathered through the browser, but feature information does not include data filled in and sent by the user. Digital browser fingerprints depend solely on the uniqueness of the device itself. Thus, even in the case of encrypted, anonymous, mobile, and even cross-domain networks, it can be identified. Modern browsers are very complex, with each component having different characteristics, and these characteristics combine to form a unique browser fingerprint.

The ultimate goal of browser fingerprinting is to track a unique user entity. Therefore, obtaining high-entropy, persistent, and preferably cross-browser characteristics of digital fingerprints is the main direction of many research efforts in the field of browser fingerprinting. The table below (see Table 1) summarizes the characteristics of various methods for obtaining digital fingerprints in a browser.

Table 1. Characteristics of methods for obtaining digital prints in a browser

Method of obtaining	Peculiarities
JavaScript-based Fingerprint	Utilizes JS functions to call various APIs for information gathering
CSS-based Fingerprint	Can be used to parse CSS strings if the user has disabled JS usage
Canvas-based Fingerprint	Characterized by high entropy and great reliability
Hardware Survey Fingerprint	Most of these algorithms require actions to be taken with measurable results, which are used as the fingerprint
Audio API Survey Fingerprint	Almost identical to the previous point
Browser Plugins Analysis Fingerprint	Involves analyzing installed browser extensions, with the list effectively serving as the fingerprint
Other Methods	Often these are exotic methods based on measuring the performance of different tasks on the device and in the network

Using a similar set of identifying features, it is possible to compile an approximate device fingerprint, as reflected in the following table (see Table 2).

Table 2. Approximate (Incomplete) Contents of a Browser's Digital Fingerprint

User-Agent	Mozilla/5.0 (Windows NT 10.0; Win64; ×64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/92.0.4495.0 Safari/537.36
Accept	text/html, application/xhtml+xml, application/xml; q = 0:9, image/webp; q = 0:8
Accept-language	ru-RU, en-US; q = 0:5
IP	84.205.22.116
Platform	Windows
Vendor	Google Inc.
appName	Netscape
Product	Gecko
cookieEnabled	True
Local time	Mon Feb 03 2024 09:57:32 GMT+0300
Resolution	1920 × 1080
WebGL vendor	Google Inc.
...	...

The digital fingerprint of a browser is complex, with different components having varying probabilities of occurrence. The uncertainty of this information can be described and measured using standard methods, the most obvious of which is the use of Shannon entropy, as suggested by the "founder" of research on browser digital fingerprinting in his work [27]. The author proposes the following line of thought: suppose there exists an algorithm for forming a browser's digital fingerprint, $F(x)$, which is analogous to a hash function. For each input browser fingerprint x, there exists a unique result from the mechanism's operation $F(x) = f_n, n \in [0, 1, \cdots, N]$.

The probability of occurrence of each result f_n is equal to $P(f_n)$. Subsequently, one can obtain the corresponding measure of uncertainty or self-information:

$$I(F(x) = f_n) = -log_2(P(f_n))$$

Next, we can round the result up to determine how many bits of information are needed to represent the information. Subsequently, the information entropy of the distribution of browser digital fingerprints is determined by the following formula:

$$H(f) = -\sum P(f_n) log_2(P(f_n))$$

Since the digital fingerprint of a browser consists of both hardware and software components, we can similarly define methods for identifying an individual browser component, for which the following will hold:

$$H_s(f_s) = -\sum P(f_{n,s}) \times log_2(P(f_{n,s}))$$

If browser components are independent of each other, their information entropy can be linearly increased, but this is often not the case. Components in most browsers are usually interconnected. For example, the Edge browser is more commonly associated with the Windows system, while the Safari browser is often tied to the iOS system. Therefore, it is appropriate to use conditional distributions for some components. Thus, the uncertainty can be expressed as follows:

$$I_{s+t}(f_{n,s}, f_{n,t}) = -log_2(P(f_{n,s}|f_{n,t}))$$

Within the scope of this article, it would be highly beneficial to consider a time-based characteristic of differences called discrepancy decay. Discrepancy decay refers to the probability that an object will change the value of attribute s over time Δt. This probability is denoted by $d^{\neq}(s, \Delta t)$.

Many events can lead to changes in a browser's digital fingerprint. In the case of the described and most logical algorithm, these events include, for example, updating the browser, updating a plugin, disabling cookies, installing a new font or an external application that includes fonts, connecting an external monitor that changes screen resolution, and so on. The author of the study [9] notes that among the users of the Panopticlick site database, the frequency of digital

fingerprint changes among returning users who accept cookies was very high: 37.4% of users visited the site more than once, and over time (within a 2–3 week period) they acquired a new digital fingerprint for their browser. By applying certain heuristic algorithms, it is quite possible to accurately identify minor changes in the browser configuration or the used device and find in the database the "previous version" of the digital fingerprint [27,29], thus ensuring device/browser recognition up to 99.5%, according to developers' claims.

5 Local Differential Privacy

The concept of Local Differential Privacy (LDP) [25,26] was proposed relatively recently as a stricter method within the differential privacy framework for preserving the confidentiality of personal information. LDP is based on the assumption that the data collector or central server cannot be trusted. In this context, each participant in the information process locally applies a differential privacy mechanism to their original data, leading to the creation of a modified version of these data. These modified data are then transmitted to an untrusted server, which acts as a collector or aggregator of data. Utilizing the modified data from all participants, the server calculates predefined statistics and publishes the statistical results. With its advantages, LDP was extensively researched and applied in industry soon after it was proposed.

There are several modern applications for the practical deployment of LDP for private statistical data collection, so LDP is used by hundreds of millions of people every day. For instance, Google's RAPPOR system [13] is based on the principle of random response and Bloom filter to implement a privacy-preserving data collection setting for mass Chrome users. This is the first and purely client-side practical solution for privacy protection that no longer requires the participation of a trusted third party and takes control of user data into its own hands. Apple announced [14] that LDP had been deployed in the iOS system, as documented in a patent application and research paper. The method is based on LDP with other techniques, including the use of Fourier transform to filter information and special subsampling methods to reduce data dimensionality.

The collection of telemetry data within the LDP framework was implemented by Microsoft in Windows 10, which uses histograms to collect data about systems and applications over time [15]. Additionally, Samsung [16] has also explored LDP technology, but it remains unclear whether it has been applied in practice.

Thus, one of the still promising and popular approaches is not to collect raw data at all. This approach was first used by Google for collecting statistics in the Chrome browser. Their system, named RAPPOR, is based on the implementation of the famous Warner's randomized response technique [17]. The randomized response works as follows: when a user wants to send a potentially sensitive piece of information to the server (for example, the answer to the question "Do you use Chrome?"), a virtual coin is "flipped" first, and if the coin lands on "heads," a random answer is returned - calculated by "flipping" another coin. Otherwise, an honest answer is sent.

The server collects responses from the entire sample of users and (knowing the probability with which the coin lands on "heads") adjusts to the existing level of "noise" to calculate an approximate answer for the truthful response. Intuitively, the randomized response protects the privacy of individual user reports because an answer "yes" could mean either "Yes, I use Chrome" or simply be the result of random coin tosses. Formally, the randomized response does indeed provide differential privacy, with specific guarantees that can be adjusted by regulating the characteristics of both the coin itself and the methods of "flipping".

Google, in its RAPPOR system, applies this relatively old technique and transforms it into something more powerful. Instead of simply answering one question, the system can compile a report on a complex vector of questions and even return complex answers, such as strings - for example, what homepage is set in your browser. This is achieved by first passing the string through a Bloom filter - a sequence of bits generated using hash functions in a very specific way. The resulting bits are then mixed with noise and summed up, with the answers being reconstructed through a rather complex decoding process. Unverified information suggests that Apple, in turn, describes and utilizes differential privacy as "using hashing, subsampling, and noise addition to enable crowdsourced learning while keeping individual user data completely private," reminiscent of the previous approach.

Furthermore, after such complex manipulations on the server side, each user is included in a certain cohort, and all standard statistics are performed exclusively by cohorts, reliably masking the data of a specific user among a mass of others. This approach closely resembles the methods used in the concept of k-anonymity.

There have been several reviews of existing ideas and technologies in the field of LDP that have emerged over the past few years. However, the rapid development of new approaches and the emergence of fresh ideas require constant attention to this topic. Similarly, and at the same speed, new possibilities related to the use of previously unconsidered technologies in connection with LDP, such as cryptography, emerge. The main mechanisms are described in the table (see Table 3).

Table 3. Mechanisms for altering or adding noise to data under LDP conditions

Mechanism	Advantages	Shortcomings	Mistake
Randomized response	Good scalability, low data processing and transmission costs	Only categorical data	$O(1)$
Laplace	Good scalability, low data processing costs (limited to noise calculation only)	Only numerical data, limited applicability	$O(\sqrt{N})$
Distortion-based	Dosing the size of information loss	Only for low-dimensional data, additional costs for "noising" data	$O(1)$
Compression-based	Control over the extent of information loss	High computational complexity for data compression and transmission, additional costs for data compression and "noising"	$O(1)$

6 Information Leaks

Research in establishing metrics applicable to the use of LDP is highly relevant. The foundations of modern approaches to accounting for privacy leaks were developed in a series of works [30,31], and subsequent research in various contexts using many different measures to ensure confidentiality. For example, differential privacy was defined in its centralized version in the context of database confidentiality to guarantee that the participation of no single individual could be disclosed based on the output of a function computed on the database (i.e., ensuring membership confidentiality). Later, in the case of differential privacy in its local version, the idea evolved into a concept where, broadly speaking, the goal is to provide equal plausible deniability for all possible values of a random variable. Concurrently with these developments, a wide range of confidentiality measures aimed at measuring the dependency between two random variables X (input) and Y (output) was proposed and studied in information theory literature. Up to now, and still, the prevailing approach to solving privacy problems has been to start with a specific definition of privacy, explore all the circumstances stemming from this definition, and attempt to develop and iteratively optimize mechanisms that guarantee a certain level of privacy and associated utility. Alternatively, one could start by defining a threat model in which an adversary with explicitly described capabilities pursues a specific goal and examines the system's vulnerability as a result of this model. That is, it involves narrowing the field of study, which suggests a significant easing of work for researchers.

The advantages of this approach can be described as follows:

- We can clearly formulate our assumptions about the capabilities of adversaries;
- We can define the concept of privacy in accordance with the threats;

– It becomes possible to develop measures to ensure privacy and objectively discuss them.

Intuitively, it may be understood that a leak is the volume of information obtained by an adversary as a result of observing the outcomes of a mechanism's operation. Following the assumptions of the authors [30], it seems natural to define this as the difference between the uncertainty regarding the input data before observing the output data and the remaining uncertainty afterwards:

$$Information\ Leak = Initial\ Uncertainty - Remaining\ Uncertainty \quad (1)$$

The question arises as to how to measure information and what can be considered as uncertainty. In relation to Eq. 1, entropy is classically measured in bits, and the initial uncertainty is defined as the minimum entropy of $X, H\infty(X)$, for example, Rényi entropy [32], obtained when $\alpha = 1$. As for the uncertainty remaining after observation, it would be natural to use the conditional minimum entropy of X given Y. Let's make some necessary definitions:

Assume all random variables are discrete and finite. Capital letters will be used to denote random variables, for example, X. Sets are denoted by uppercase letters, for instance, χ represents the alphabet of X. We use P_X to represent the probability distribution of X, and $supp(P_X) := x \in X : P_X(x) > 0$ to represent the support of P_X. If n is a positive integer, then we use $[n] := 1, \cdots, n$ to describe the set of all natural numbers ranging from 1 to n. Additionally, $log(\cdot)$ will be used to denote the natural logarithm.

For random variables X and Y, we will use $P_{Y|X}$ to denote the conditional probability of Y given the random variable X. Throughout this article, we use X to represent some data containing confidential information (i.e., data of a single user, statistics, or some secret symbol), the random variable $P_{Y|X}$ to represent the private mechanism responding to queries about X, and Y to represent the responses to public queries.

Next, we may be interested in quantitatively assessing the leakage under a given protocol, i.e., the amount of information about the input data that an adversary can learn by executing the protocol and observing the resulting output data.

Before launching the private protocol, the probability of correctly guessing the value of x depends only on the prior distribution of X, and therefore, a conventional rational adversary would assume that the most probable value of x, called the prior probability of a correct guess, would be the value of x that has the maximum prior probability $PR(X) = max_x(P_x)$.

Many authors, based on definitions by Smith [31], propose the following definition of conditional entropy:

$$H_\infty(X) = -log(PR_o(X))$$

Thus, based on this definition, one can define leakage as follows:

$$L(X;Y) = H_\infty(X) - H_\infty(X|Y) =$$
$$-log(PR_o(X)) + log(PR(X)) = log(PR(X)/PR_o(X))$$

Interestingly, the greatest leakage occurs in the case of a uniform distribution [30].

Many studies in this field have expanded the model described above [33–36], using the concept of leakage $L(X;Y)$ as a tool for quantitatively assessing privacy. The results of these studies suggest that the maximum leakage calculated in this way is a convenient measure of privacy with a reliable definition, which can be used in a variety of different scenarios. However, one of the apparent limitations of such a definition is that the maximum leakage is determined for the average result Y. Therefore, such privacy guarantees do not provide the ability to distinguish between individual possibilities of leaks based on their outcomes.

7 Impact of Features on Privacy Leakage

Within the framework of this article, approaches to assessing privacy leakage associated with identifying features of digital device fingerprints are considered. Utilizing a dataset, each element of which includes identifying features listed in Table 2, the following assessment algorithm was developed:

1. **Definition of input data**: the dataset D consists of n elements, each incorporating identifying features F_1, F_2, ..., F_m.
2. **Calculation of entropy for each feature**: for each feature F_i, calculate its entropy $H(F_i)$ as a measure of uncertainty or diversity of feature values.

Algorithm 1. Calculation of entropy for each feature

```
function CalculateEntropy(F)
count ← map of (value → frequency)
for v in F do
    count[v] ← count[v] + 1
end for
total ← length of F
H_F ← 0
for each (value, frequency) in count do
    p ← frequency/total
    H_F ← H_F − p · log2(p)
end for
return H_F
```

3. **Analysis of mutual information**: for each pair of features F_i and F_j, calculate the mutual information $I(F_i;F_j)$ to assess the degree of their interdependence.

Algorithm 2. Analysis of mutual information

```
function CalculateMutualInformation(F, G)
H_F ← CalculateEntropy(F)
H_G ← CalculateEntropy(G)
joint_count ← map of ((f, g) → frequency)
for each pair (f, g) in zip(F, G) do
    joint_count[(f, g)] ← joint_count[(f, g)] + 1
end for
total ← length of F
H_FG ← 0
for each ((f, g), frequency) in joint_count do
    p_fg ← frequency/total
    H_FG ← H_FG − p_fg · log_2(p_fg)
end for
return H_F + H_G − H_FG
```

4. **Data aggregation**: using the obtained measures of entropy and mutual information, aggregate the data to obtain an overall indicator of privacy leakage through compositional information security measures.

Algorithm 3. Data aggregation

```
function AggregateData(measures)
result ← 0
for measure in measures do
    result ← result + measure
end for
return result
```

5. **Correlation analysis**: determine the correlation of privacy leakage relative to the number of identifying features in the dataset.

Algorithm 4. Correlation analysis

```
function CorrelationAnalysis(X, Y)
mean_x ← Average(X)
mean_y ← Average(Y)
numerator ← 0
denominator_x ← 0
denominator_y ← 0
for Each(x, y) in Zip(X, Y) do
    numerator ← numerator + (x − mean_x) * (y − mean_y)
    denominator_x ← denominator_x + (x − mean_x)^2
    denominator_y ← denominator_y + (x − mean_y)^2
end for
return numerator/√(denominator_x * denominator_y)
```

In this study, the input data were collected from 15 different, non-repeating devices using a variety of browsers. These initial data included unique digital fingerprints of each browser, incorporating information such as operating system version, installed fonts, screen resolution, and other browser data that could serve as identifying features.

To expand the dataset and increase its diversity, the initial data were artificially modified and augmented, allowing the generation of new variant digital fingerprints. As a result of this process, the total volume of data increased to 100 different digital fingerprints. Special attention was given to ensuring a high level of entropy during data generation to maximize uncertainty and complexity, which is critically important for privacy analysis.

Based on this enhanced sample, a study was conducted aimed at averaging privacy loss and analyzing the impact of the number of identifying features on the level of data leakage. The results were presented in a graph (see Fig. 1) that illustrates the nonlinear increase in privacy leakage as the number of features in the dataset grows. This analysis underscores the importance of careful data control and protection when working with large volumes of information and multiple identifying characteristics.

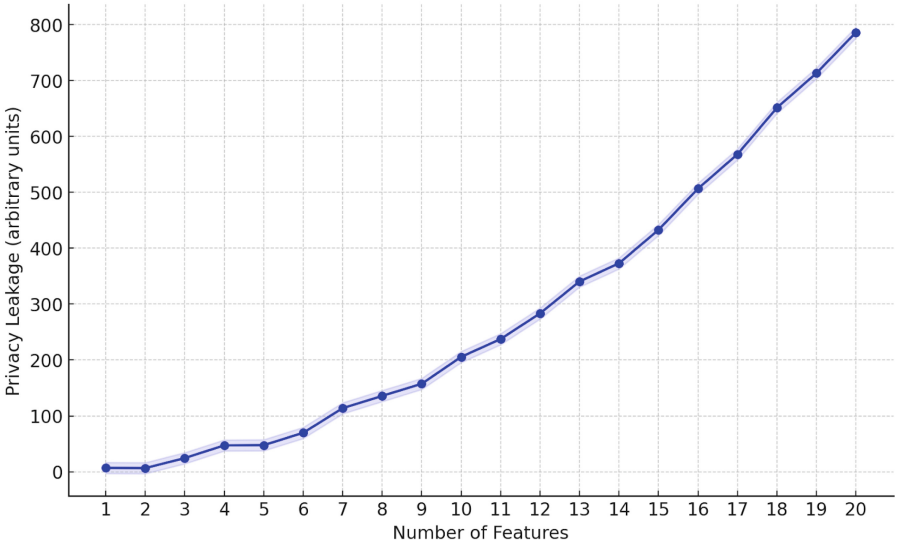

Fig. 1. Dependency of Privacy Leaking on Number of Features

In this case, privacy leakage increases quadratically as the number of features increases. This is because adding each new feature not only linearly increases the amount of information but also enhances the complexity of the interrelationships and combinations among features, leading to a more significant growth in

potential leaks. Such a conclusion underscores the need for stricter confidentiality measures when working with large datasets.

The graph displays deviations from the norm (highlighted in light blue), illustrating possible variations in privacy leakage due to random factors or changes in the data. This deviation could represent differences in leakage due to the varying effectiveness of implemented security measures or the specifics of data processing under different conditions.

8 Minimization Models

We may be interested in the fact that in many cases, a digital fingerprint is a kind of hash, the value of which is unstable over time and computed by various algorithms. Thus, a digital fingerprint can also be considered a random variable. Therefore, willingly or unwillingly, in the presence of a digital fingerprint, there may exist multiple databases or database fragments containing tuples with the digital fingerprint as a user identifier. These records can be formed either using Local Differential Privacy (LDP) technologies or without this concept. In this case, any fragment can be considered auxiliary (for the adversary) to the others (see Fig. 2).

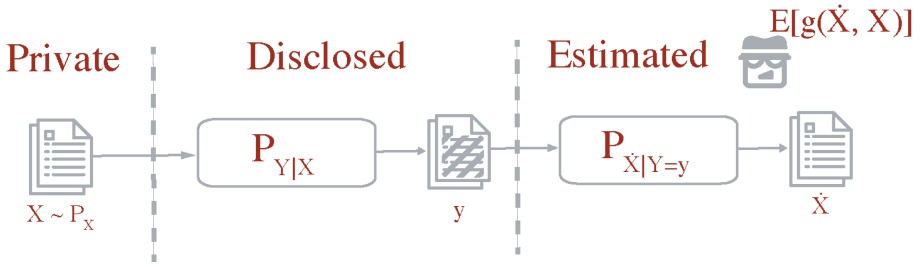

Fig. 2. Information Leakage Minimization Model

One can conceive an information leakage minimization model. Suppose we are given a random variable X with distribution PX, and by using the mechanism PY—X, we can add noise to these sensitive data. It's clear that in this case, it would be quite difficult for an adversary to construct a new estimate. The function's result will depend on the balance between data privacy and utility. In some instances, a compromise between these two aspects may be required.

Recalling the thesis about the utility of a DP algorithm degrading to zero when an adversary possesses external or auxiliary information, one might assume that in conditions of widespread use of digital fingerprints, it is necessary to move from classical definitions of differential privacy to some new definitions of privacy and utility that would allow for the evaluation of mechanism privacy. In doing so, of course, an attempt should be made to define the relationships of new algorithms with those algorithms already practically recognized as classic (see Fig. 3).

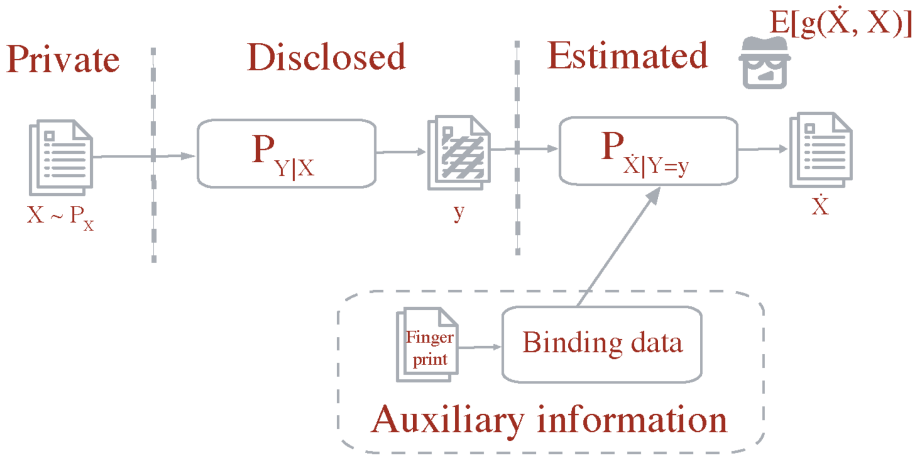

Fig. 3. Information Leakage Minimization Model Using Auxiliary Information

Let's assume X is a random variable representing some data containing confidential information. We'll keep X abstract to maintain generality. X could be a single data record obtained from an individual (local version of DP), X might represent an entire database containing confidential information (centralized DP), or X could be a secret symbol, such as a password (under potential side-channel attack conditions).

Next, suppose Y is the output of a randomized function with input X. For example, in the context of local DP, Y denotes the noised data of a single user, collected by an untrusted data server. In centralized DP conditions, Y might be some statistics (sum, average, etc.), computed from the database, and in side-channel conditions, Y could be some physical output data related to the input X, like delays between keypresses when entering a password. In all these scenarios, we are interested in answering the following question: how much information do we gain about X through observing Y?

9 Conclusion

This study focuses on exploring the capabilities of privacy leakage assessment in the context of an untrusted client-server dialogue within the framework of Local Differential Privacy (LDP). In the realm of collecting anonymous statistical data and utilizing device fingerprints, the issue of privacy becomes critically important.

The research is highly relevant as there is a need to process data in conditions of distrust between the client and the server, as well as to maintain the privacy and anonymity of users. The application of the LDP methodology allows for the minimization of potential privacy loss during data collection and analysis. This research contributes to the development of entropy assessment algorithms that

can be used for data aggregation and statistical indicator extraction without disclosing personal information.

The findings of this research can be applied in various fields, such as user behaviour analysis, big data processing, and decision-making based on statistical data. It is anticipated to explore algorithms and methods for ensuring privacy that can be implemented in real systems, thereby adhering to regulatory requirements and increasing user trust in the process of processing and using their data.

Research in this direction can make a contribution to the fields of data privacy and the processing of anonymous statistical data under conditions of untrusted client-server dialogue. The algorithms and privacy assessment methods developed within the LDP framework can provide a balance between the use of sensitive data and the protection of user privacy, contributing to the development of safe and trustworthy information processing in the modern information society.

References

1. General Data Protection Regulation (2016). https://gdprinfo.eu/
2. Brazilian General Data Protection Law (2019). https://iapp.org/media/pdf/resource_center/Brazilian_General_Data_Protection_Law.pdf
3. The Digital Personal Data Protection Bill (2022). https://www.meity.gov.in/writereaddata/files/The%20Digital%20Personal%20Data%20Protection%20Bill%2C%202022.pdf
4. Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '03, pp. 202–210 (2003)
5. Sweeney, L.: k-ANONYMITY: a model for protecting privacy. Internat. J. Uncertain. Fuzziness Knowl. Based Syst. **10**(5), 557–570 (2002)
6. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: L-diversity: privacy beyond k-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering (ICDE), p. 24, Atlanta, GA, USA, April 2006
7. Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: privacy beyond k-Anonymity and l-diversity. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering (ICDE), pp. 106–115, Istanbul, T:urkey, April 2007
8. Facebook-Cambridge Analytica data scandal (2018). https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal
9. How to Protect Yourself From the Giant Marriott Hack (2018). https://www.wired.com/story/marriott-hack-protect-yourself/
10. The Top 12 Data Breaches of 2019. https://www.securitymagazine.com/articles/91366-the-top-12-data-breaches-of-2019/
11. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately. SIAM J. Comput. **40**(3), 793–826 (2011)
12. Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Local privacy and statistical minimax rates. In: Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS), pp. 429–438, Berkeley, CA, USA, October 2013
13. Erlingsson, U., Pihur, V., Korolova, A.: RAPPOR: randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 1054–1067, Scottsdale, AZ, USA, November 2014

14. Differential Privacy Team Apple: Learning with privacy at scale. Apple Mach. Learn. J. **1**(8), 1–25 (2017)
15. Ding, B., Kulkarni, J., Yekhanin, S.: Collecting telemetry data privately. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), pp. 3572–3581, Long Beach, CA, USA, December 2017
16. Nguyen, T.T., Xiao, X., Yang, Y., Hui, S.C., Shin, H., Shin, J.: Collecting and analyzing data from smart device users with local differential privacy (2016). https://arxiv.org/abs/1606.05053
17. Warner, S.L.: Randomized response: a survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. **60**(309), 63–69 (1965)
18. Braun, C., Chatzikokolakis, K., Palamidessi, C.: Quantitative notions of leakage for one-try attacks. Electron. Notes Theor. Comput. Sci. **249**, 75–91 (2009)
19. Wang, N., Xiao, X., Yang, Y., et al.: Collecting and analyzing multidimensional data with local differential privacy. In: Proceedings of the 2019 IEEE 35th International Conference on Data Engineering (ICDE), pp. 638–649, Macao, China (2019)
20. Ren, X., Yu, C.-M., Yu, W., et al.: LoPub: high-dimensional crowdsourced data publication with local differential privacy. IEEE Trans. Inf. Forensics Secur. **13**(9), 2151–2166 (2018)
21. Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D., Wang, T.: Privacy at scale: local differential privacy in practice. In: Proceedings of the 2018 International Conference on Management of Data (SIGMOD), pp. 1655–1658, Houston, TX, USA, June 2018
22. Li, N., Ye, Q.: Mobile data collection and analysis with local differential privacy. In: Proceedings of the 2019 20th IEEE International Conference on Mobile Data Management (MDM), pp. 4–7, Hong Kong, June 2019
23. Kifer, D., Machanavajjhala, A.: No free lunch in data privacy. In: SIGMOD (2011)
24. Liu, C., Chakraborty, S., Mittal, P.: Dependence makes you vulnberable: differential privacy under dependent tuples. In: Network and Distributed System Security Symposium (NDSS), pp. 21–24 (2016)
25. Dwork, C.: Differential privacy. In: Automata, Languages and Programming, pp. 1–12. Springer, Heidelberg (2006). ISBN 978-3-540-35908-1
26. Dwork, C., Naor, M.: On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. J. Privacy Confidentiality **2**(1) (2010)
27. Eckersley, P.: How Unique Is Your Web Browser? Electronic Frontier Foundation (2010)
28. Bessonova, E.: Method of identifying users on the Internet using a component profile. Dissertation for the degree of candidate of technical sciences, ITMO, p. 28 (2012)
29. Fingerprint Platform. https://dev.fingerprint.com
30. Braun, C., Chatzikokolakis, K., Palamidessi, C.: Quantitative notions of leakage for one-try attacks. Electronic Notes in Theoretical Computer Science, pp. 75–91
31. Smith, G.: On the foundations of quantitative information flow. In: Proceedings of the 12th International Conference on Foundations of Software Science and Computation Structures, Lecture Notes in Computer Science (2009)
32. Reny, A.: On measures of entropy and information. In: Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics, and Probability (1960)
33. Espinoza, B., Smith, G.: Min-entropy as a resource. Inf. Comput. **226**, 57–75 (2013)
34. Alvim, M.S., Chatzikokolakis, K., Palamidessi, C., Smith, G.: Measuring information leakage using generalized gain functions. In: 2012 IEEE 25th Computer Security Foundations Symposium, pp. 265–279 (2012)

35. Alvim, M.S., Chatzikokolakis, K., Mciver, A., Morgan, C., Palamidessi, C., Smith, G.: Additive and multiplicative notions of leakage, and their capacities. In: 2014 IEEE 27th Computer Security Foundations Symposium, pp. 308–322 (2014)
36. Alvim, M.S., Chatzikokolakis, K., McIver, A., Morgan, C., Palamidessi, C., Smith, G.: The Science of Quantitative Information Flow. Springer (2020)
37. Podsevalov, I., Podsevalov, A., Korkhov, V.: Differential privacy for statistical data of educational institutions. In: Computational Science and Its Applications - ICCSA 2022 Workshops, pp. 603–615 (2022)