K-Anonymity Versus PSI3 for Depersonalization and Security Assessment of Large Data Structures

Alexander Bogdanov[1,2(✉)], Nadezhda Shchegoleva[1,2], Valery Khvatov[3], Gennady Dik[1,4], Jasur Kiyamov[1], and Aleksandr Dik[1]

[1] St. Petersburg University, Saint Petersburg, Russia
{a.v.bogdanov,n.shchegoleva,z.kiyamov,a.dik}@spbu.ru
[2] St. Petersburg State Marine Technical University, Saint Petersburg, Russia
[3] DGT Technologies AG, Ottawa, Canada
[4] St. Petersburg LLC "System Technologies", St. Petersburg, Russia
g.dick@systechnologies.ru

Abstract. The article presents a modern approach to protecting personal data from de-anonymization. Descriptions of the k-anonymity model using entropy, with one-way publication of information, and the PSI3 technique using homomorphic encryption, with bilateral or multilateral information exchange, examples of the use of metrics for assessing the quality of depersonalization and their improvement, recommendations for using methods in the current activities of personal data operators are given.

Keywords: information protection · personal data · k-anonymity · entropy · PSI3

1 Introduction

Information systems supports the big data ecosystem process personal data everywhere. Processing of personal data is a set of actions (operations) performed with or without the use of automation tools with personal data, including collection, recording, systematization, accumulation, storage, clarification (updating, changing), extraction, use, transfer (distribution, provision, access), depersonalization, blocking, deletion and destruction.

Protection of personal data is a measures set of technical, organizational and organizational-technical nature aimed at protecting information related to a certain or determined on basis of an individual information (subject of personal data). As with any other information, personal data security breaches falls into three categories:

1. confidentiality (inaccessibility of information to be available to unauthorized persons);

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14109, pp. 317–333, 2023.
https://doi.org/10.1007/978-3-031-37120-2_21

2. integrity (preservation of completeness and consistency of information);
3. availability (information is available and ready for use by authorized participants in the information exchange).

There are few of key examples of personal data protection:

- Individuals have the rights to control how their personal data is used. Operators must obtain explicit and informed consent from individuals before collecting, using or disclosing their personal data;
- Operators must only collect personal data for specific, legitimate purposes and do not use it for other purposes without obtaining the consent of the individual;
- Operators must only collect and process the minimal amount of personal data necessary to achieve their legitimate purposes;
- Personal data operators must take reasonable steps to ensure that personal data is accurate, complete and up-to-date;
- Operators must implement appropriate technical and organizational measures to protect personal data from unauthorized or unlawful processing of accidental loss, destruction or damage;
- Individuals have certain rights in relation for their personal data, including the right to access, correct, delete and restrict the processing of their personal data;
- Operators are responsible for complying with applicable data protection laws and regulations, taking appropriate steps to ensure that their employees and third party service providers also comply with laws and regulations.

At the same time, an important point in risk assessment is not the fact of gaining access to an array of anonymized data (developers, system administrators have a significant amount of access to data), but the fact that attack is completed. If the organization has mechanisms for detecting an attack and preventing it before the attack is completed, i.e. the attack began, but did not end and reach the goal, then the risk was not realized, since the result - depersonalization - was not achieved. Such mechanisms are included in the technical and administrateve aspects of the organization. To determine contextual risks, this criterion of completeness is important - the data must be de-anonymized, which creates the possibility of causing damage to the personal data (PD) subject.

In general, the specific measures required to protect different types of personal data will be depend of the context and information confidentiality [1]. In general, the protection of personal data requires an integrated approach that includes strict access control, encryption, authentication measures, and strict protocols for data processing (anonymization) and storage. In addition, organizations must comply with applicable laws and regulations, regarding the collecting, using and disclosuring of personal data.

This article continues and improves the approach described in [2], based on calculated value of the re-identification risk.

The risk approach to personal data security is a proactive and systematic approach that enables organizations to continuously assess and manage the privacy

risks associated with personal data. By following this approach, organizations can ensure that their customers personal data is protected, employees and other stakeholders can be confident that their personal data is handled.

2 Modern Approach to Data Protection

PD security differs from the classical concept of information security [3,4], the focus of which is unauthorized activity (which actually leads to loss of confidentiality, integrity or availability). When processing PD for the purpose of depersonalization, we are talking about planned activities, which, however, may create consequences for the private life of individuals. PD confidentiality threats arises as a result of authorized data processing and as a result of unauthorized access (actions of an attacker):

- As a result of unauthorized access to data - emotional suffering of individuals, economic losses from identity theft, as well as physical or psychological harm due to persecution;
- As a result of a decrease in the quality of information processing (decrease in it's value) within the framework of planned processing - a decrease in the quality of services provided, failures in the operation of systems, incorrectly made or delayed decisions, which may also affect the health of individuals or entire population groups.

The key provisions of the proposed risk model depersonalization procedure take into account confidentiality issues that arises as a result of unauthorized access and planned data processing:

- The risk of publishing (exchanging) anonymized data is a measure of the extent where the subject is threatened by potential circumstances or events and remains as a probability function P of the occurrence of such events and the adverse consequences violation of the personal data confidentiality - I, Impact:
$$R = P * I \tag{1}$$
- The construction of completely anonymized data sets that store useful information is an idealized hypothesis that is not feasible in practical solutions. The proposed approach based on the likelihood of re-identification attacks and setting an acceptable risk;
- The process by identified risks, according to [4] is called risk assessment;
- Risk assessment requires a risk model that defines risk factors and the relationships between them:
  - Risk factorization (set selection of individual risk components and establishing a relationship between them);
  - Release model of data;
  - Set quantitative risk thresholds;
  - The required usefulness level of the received depersonalized data;

- Building a risk model procedure for a specific depersonalization procedure, including the possibility of reassessing the risk when using various depersonalization methods

Damage may include:

– Compliance costs associated with privacy concerns for individuals;
– Direct costs of fines and legal fees;
– Business loss, refusal of using products or services;
– Reputational costs leading to loss of trust from users;
– Decreased performance or inability to achieve the organization's mission.

Specific damage can be assessed for each specific situation; within the framework of this model, proposed to confine ourselves vector of values:

$$I = [MinorDamage = I_1; MediumDamage = I_2; HighDamage = I_3] \quad (2)$$

Thus, the risk assessment proposed in this model contains damage as a parameter; in what follows that risk assessment is understood as an assessment probability of occurrence events $P \in [0; 1]$.

The general factorization formula is the probability of the re-identification risk:

$$P_{re-id} = P_{context} * P_{data} \quad (3)$$

Here: P_{re-id} - general risk of re-identification, $P_{context}$ - contextual risk, defined by a set of organizational and technical risks, P_{data} - data risks

Data risks are associated with a specific data set and should be calculated taking into account the characteristics of the data set - in particular, the allocation of direct identifiers and quasi-identifiers. Direct identifiers and their substitution are the smallest problem in solving the reidentification risk problem. Correct quasi-identifiers identification and depersonalization application methods taking into account their structure and type, is a much more difficult task, bearing in mind the probabilistic approach to risk assessment.

Transformation of PD (depersonalization), taking into account the risk model, requires a balance between the usefulness data obtained as a result of depersonalization and the acceptable level of risk. The acceptable threshold for the risk of re-identification personal data depends on various factors, including the data confidentiality, context of its use, applicable laws and regulations. Generally, the risk threshold for re-identification is considered being very low, since any re-identification of personal data may lead to a breach of confidentiality and security. Thresholds risk is setting according to use cases.

The acceptable risk threshold quantification for the personal data re-identification might vary depending of the specific context and requirements for data use. In some cases, a certain percentage value, such as 3%, 10%, or 20%, may be used to determine threshold. These percentage meanings represent the acceptable probability of re-identifying individual from the dataset.

For example, some countries regulations requires that personal data must be anonymized in the point where the risk of re-identification is below a certain

percentage, such as 3% or 5%. In other cases, a threshold may be set based on the data confidentiality and the potential harm that may result breach of confidentiality. A general idea orders of the risk can be drawn from the Table 1 below [5]:

Table 1. Applying k-Anonymity

Impact on PD privacy	Acceptable vice of risk	Equivalence class size for aggregated data
Low	0.1	10
Medium	0.075	15
High	0.05	20

The acceptable level of risk is determined taking into account the following parameters:

- Likelihood of a threat to confidential data ($P_{context} * P_{data}$);
- Potential damage I (assessed in terms of "low", "medium", "high" or through points) - see [6,7]. The damage is also determined taking into account the volume of published data and the sensitivity of the attributes;
- Publication scenarios (data dissemination scenarios) - influences the recommended equivalence classes and therefore data risk;
- The business model of the organization publishing the data ("risk appetite"). This parameter describes risk tolerance taking into account the characteristics of the business (industry), the value of assets (data containing private characteristics), management preferences;

3 Risk Assessment Metrics

Although, calculation metrics risk may be differ by depending of used methods. Here are general approaches that allow quantifying the probability based on data behavior (homogeneity, difference between the original data and ones anonymized, etc.).

The metrics are calculated in consideration with the attributes (fields) of the sets, which classification is shown in the Fig. 1 below [6].

According with accepted approach [7], the greatest influence on metrics is exert by quasi-identifiers, which combination makes possible to indirectly determine the individual and compare his sensitive attributes. The main idea of risk calculation which based on attribute selection is to form a group of similar records (equivalence classes), where impossible to match only one individual and his sensitive attributes. Then probability of re-identification is directly related with the size of equivalence classes (inversely proportional to equivalence class), and also depends of distribution of sensitive attributes in such equivalence classes. The Table 2 provides summary information of the various metrics with specified subject dependencies.

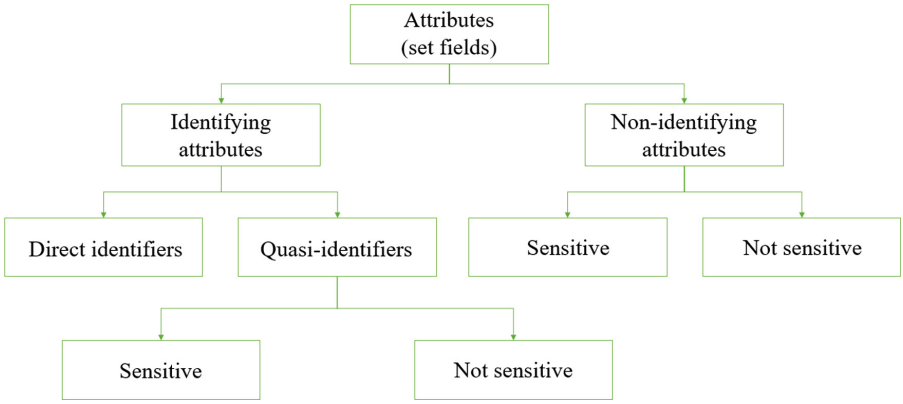


Fig. 1. Classification of attributes

Table 2. Evaluation of entropy by attributes for the choice of quasi-identifiers

№	Metrics	Explanation and calculation method
1	k-Map	Measures degrees anonymity achieved by anonymized dataset, based on smallest size of records group that share same Quasi-Identifier (QI) values. In other words, the metric measures have minimal number of records, needed for attacker to re-identify a person in a dataset.
2	k-Anonymity	Anonymity degree measure in dataset. Dataset is called to be k-anonymous if each entry in dataset is indistinguishable from at least other k-1 entries.
3	L-Diversity	A measure of sensitive diversity attributes values in a group of k-anonymous records. It measures probability that an attacker will be able to re-identify a person based on quasi-identifiers combination and sensitive attributes.
4	t-Closeness	A measure of similarity between distribution of sensitive attribute values in the original dataset and anonymous dataset. This measure extends which anonymized dataset retains the statistical properties from the original dataset

The calculated metrics makes possible to evaluate numerical probabilities of re-identification risks by using formulas from the form.

$$P_{data} = \frac{1}{f} \tag{4}$$

where f - one of the metrics meanings, listed in the table above.

In accordance with these descriptions, depersonalization calculation metrics is applies for specific datasets, both determined by their attributes (columns), and directly by the records in set - the match frequency for individual attributes. For example, to calculate k-anonymity metric (i.e. ensure that each record in data sets is indistinguishable from at least another k-1 records), must be taken these following steps:

- Conduct identification of quasi-identifiers (Qi) in data set (fields or attributes in dataset that could be used to associate a person with his confidential information). For example, Qi include demographic information such as age, gender, and a zip code;
- Group the records in data set according with the key interfaces values. This creates a set of groups, in which each group contains entries that have the same values for Qi identifiers;
- Generalize or suppress Qi."Generalization" means replacing a specific value with more general value, such as replacing a date of birth with an age range."Suppression" means complete value removal;
- Make sure that each group has at least k-entries. If group contains less k-records, it is necessary to merge it with another group and generalize, or suppress Qi until the group satisfies the k-anonymity requirement;
- Repeat the previous steps until all groups meet k-anonymity requirement.

4 Practical Implementation of Risk Assessment

4.1 Using the Models K-Anonymity

When unilaterally information is publishing (for example, generating machine learning data), it's convenient to use classical depersonalization, developed by k-Anonymity model. The k-Anonymity model is based on quasi identifiers allocation in the set attributes (columns). There is practical algorithm implementation. Main method steps are include:

1. Identifier highlighting - attributes defining or their combinations that can potentially identify individuals in dataset;
2. Records grouping (finding equivalence classes) - removing direct identifiers and clustering records, bases on their quasi-identifiers values. Each group must obtain at least k-quantity entries to achieve k-anonymity (other metrics may also be used);
3. Quasi-identifiers generalization - summarizing the meanings of quasi-identifiers in each group to make them less specific and more similar. Generalization degree depends on confidentiality requirements and information loss level, which is acceptable to solve the problem;
4. Suppression of sensitive values - suppressing or removing sensitive attributes or values that may be associated with specific users in each group (suppression can also be done by randomizing values);

5. Depersonalization level evaluation - level assessment of depersonalization is determined by used metric. In this case, probability of re-identification is estimated as inversely value, proportional to metric value. Each entry in dataset has its own value equal to the size of corresponding equivalence class, depending on used offender model selected the smallest metric (highest probability), and the average metric for the entire set;
6. Checking the level of information loss - using depersonalization techniques (step 3.4) inevitably leads to information loss. Numerical value of corresponding metric gives a relative loss estimate, which acceptability is ultimately determined by testing problem application being solved during publication;
7. Reapplying Depersonalization Techniques - if assessment metrics risk (depersonalization level) does not provide the required security level (comparison with threshold risks) or information loss level is high, corresponding depersonalization parameters (the "blurring" level) could be changed. In this case, new iteration is carried out, starting from step 2;
8. Total Risk Calculation - when we considering contextual risk, overall probability is calculated by using a formula, which compares probability threshold;
9. Re-identification risks monitoring - tracking and assessing re-identification risks of given information set is adds over time as a new data, or new available becoming information.

Below is the pseudo-code implementation sample of the k-Anonymity method (Algorithm 1) according to the algorithm.

The algorithm consist next following steps:

1. Collect the data from given columns by using the group method and count the number of records in each group;
2. Determine minimal number of records in each group by using lambda function method;
3. Count the number of groups with a certain minimal number entries by using the Counter function;
4. Select 5 smallest values for minimal number of entries;
5. Find the unique number values in "Account" column;
6. Calculate k-anonymity as ratio records number in the dataset to number of unique values in "Account" column;
7. Return consisting pairs of values list (k, the percentage of groups with the minimal number of entries), where k takes the values chosen in step 4;
8. In the output, the function returns a pairs of values list representing k values, and percent the groups with the minimal number of records in the data. These values allow you to assess degree anonymity data.

The function takes two input arguments:

– data - is a data set, in which k-anonymity parameter should be calculated;
– columns - is a list of columns, in which data should be grouped.

Algorithm 1. calculating the k-anonymity metric

```
1: function K_ANONYMOUS(data, columns)
2:     aggregrate ← data.groupby(columns).count().apply(lambda x : set(x) -
   set({0}),axis = 1)
3:     aggregrate ← aggregrate.apply(lambda x : 0 if x == set() else min(x))
4:     count ← dict(Counter(aggregrate))
5:     sort_arg ← sorted(set(aggregrate))[:5]
6:     dub_data ← data["Account"].drop_duplicates()
7:     nyu ← round(len(data)/len(dub_data))
8:     print(len(dub_data))
9:     return [[arg*nyu,round(count[arg]/len(aggregrate) * 100,3)] for arg in
   sort_arg]
10: end function
```

The k-Anonymity model is a privacy preservation technique that protects to identity people in dataset by grouping them into anonymous groups, based on their quasi-identities. Quasi-identifiers are attributes (or attributes combinations) which potentially identify individual in a dataset. Algorithm 1 above is assumes a "manual" choice of quasi-identifiers, which represents as complex structure. Within of the k-anonymity model, various metrics could be used, making possible to quantify the re-identification probability. Some approaches allows automating:

- Entropy-based approach: this approach involves calculating the entropy of each attribute in the dataset and selecting features with the highest entropy values as quasi-identifiers. Attributes with high entropy have more unique values, making them more likely to distinguish between people.
- Correlation-based approach: the approach involves identifying attributes that are highly correlated with each other since they may reveal sensitive information when combined. Then, highly correlated attributes are selected as quasi-identifiers.
- Domain knowledge-based approach: this approach is similar to "hand-picked" selection, involves using domain knowledge and expert judgment to determine which attributes are likely to be quasi-identifiers. This approach can be useful in work with datasets containing sensitive or confidential information.
- L-diversity-based approach: this approach involves selecting quasi-identifiers, based on their ability to achieve l-diversity which represents as a measure of sensitive diversity attributes in each group. The approach idea lies that each group of selected attributes has a sufficient level of diversity in sensitive values, what makes difficult to identify people.
- Outlier-based approach: the approach involves identifying outliers in the dataset and using outlier-related attributes as quasi-identifiers. Within some sets, outliers are individual related records that have unique different attribute values against to other records in terms of selected attributes, making them more likely being identified.

Below is the algorithm pseudocode (Algorithm 2) that calculates the entropy [10,11] for each column of the quasi-identifier in the dataset. Algorithm consists these following steps:

- Create an empty list ent_array to store entropy values for each quasiidentifier;
- Pass through each quasi-identifier in QI (set of quasi-identifiers);
- Group the data by current quasi-identifier and count the number of occurrences;
- Calculate the total number of occurrences;
- Initiate entropy in the current quasi-identifier aims to 0;
- Pass through each count and calculate the probability contribution into entropy;
- Add the entropy value into the ent_array list;
- Print the entropy value for the current quasi-identifier.

Algorithm 2. Attribute entropy calculation

```
ent_array ← []
for each qi in QI do
    counts ← df.groupby(qi).size()
    total ← counts.sum()
end for
entropy ← 0
for each count in counts do
    probability ← count/total
    entropy ← entropy − probability ∗ log2(probability)
end for
ent_array.append(entropy)
print("Entropy of ", qi, ": ", entropy)
```

4.2 Using the PSI3 Algorithm

Some information exchange scenarios involve bilateral or multilateral data exchange in order to form a common data array (intersection set). Such data set can be formed by using the Private Set Intersection (PSI) [12] technique.

For example consider the information exchange between A organization (advertising platform) and B organization (bank, financial institution). Organization A has a list of customers with phone numbers and additional information (such as number of viewed ad times). Organization B (bank) also has a list of customers with phone numbers and banking information (who, when and how much they spent). To solve the problem you need to form total set of ad views against costs that gives advertising effectiveness (advertising touch, marketing attribution). Both organizations have a number of restrictions in phone numbers transfer, therefore using of hashing schemes is not acceptable due to significant sensitive information amount. To solve the problem of constructing confidential intersection sets (which hides phone numbers) can be used various approaches:

- Homomorphic encryption scheme [9]: an encryption scheme that allows calculations to be performed on encrypted without decrypting this data. Example of homomorphic encryption scheme lays in Payet cryptosystem;
- Invisible/forgetful transmission: this protocol allows one party send a message to another party without knowing anything about its content, so sender receives recipient message without knowing anything about his choice. Example of a forgotten transfer protocol is 1-of-2 Oblivious Transfer, as well as Bloom filters.

Confidentiality-preserving intersection of data sets can be arranged by using the PSI3 approach. PSI3 [13] (Privacy-Preserving Sequential Subgroup Summation and Intersection) is a privacy preservation algorithm for calculating subgroups intersection within larger group, providing differential privacy in combination with confidential computing mechanisms.

The PSI3 algorithm works by dividing a large group into smaller subgroups and calculating the sum of each. Then it adds Laplace noise into each of these subgroup sums to preserve differential privacy. After that, algorithm computes the subgroup intersection by comparing the sums of the noise in subgroups. Finally, it adds more Laplace noise to intersection result for better protection privacy of the people in subgroups.

Key advantage of PSI3 algorithm is about allowing intersection queries to be computed over subgroups without exposing any individual level information. This makes it useful for scenarios with issue privacy, such as healthcare, finance, or government data sharing. Although, the PSI3 protocol can be used in industrial implementations, there are some use restrictions:

- Computational complexity: the PSI3 algorithm involves several rounds of communication and computation, which can be computationally intensive and consuming time. This can be especially tricky when dealing with large datasets.
- Difficulty in communication: the PSI3 algorithm includes multiple rounds of communication between the parties, which can be a significant barrier in situations where the parties are geographically dispersed or communication channels are unreliable.
- Errors accumulation: each PSI3 algorithm round introduces some error degree into computation of subgroup intersections. This error can accumulate over several rounds. Care must be taken to ensure that error lies within acceptable limits.
- Sensitivity to outliers: the PSI3 algorithm assumes that the data is distributed according to a normal distribution, which may not always be in the case. In particular, the algorithm may be sensitive to outliers in the data, which may affect the accuracy of the results.
- The need for trust: the PSI3 algorithm requires parties to trust each other to follow the protocol, and not deviate from agreed procedures. This can be challenging in situations where is no trust, or where the parties have competing interests.

Below generalized implementation method in pseudocode (Algorithm 3).

Algorithm 3. PSI3 algorithm

```
public_key, private_key ← paillier.generate_pailier_keypair()
function ENCRYPT-RECORDS(records)
    encrypted-record ← public_key.encrypt(records)
    return encrypted-record
end function
function INTERSECT
    bank_records ← bank.records
    ad_records ← advertisement_hub.records
    result ← []
    for each bank_record in bank_records do
        for each ad_record in ad_records do
            if bank_record[0] ← ad_record[0] then
                result.append((encrypt_records(bank_record[0])
            end if
        end for
    end for
    return bank_records
    return risk_level
end function
function CALCULATE-RISK(epsilon)
    intersection_size ← len(intersect())
    bank_size ← len(bank.records)
    advertisement_hub_size ← len(advertisement_hub.records)
    privacy_budget ← bank_size * advertisement_hub_size / intersection_size
    risk_level ← ϵ / privacy_budget
    return risk_level
end function
```

The algorithm describes PSI3 class, which is using to analyze data privacy risks in two data sources: bank records and advertising hub. The class includes these following methods:

- constructor PSI3 (bank, advertisemen_hub, public_key): this method initializes PSI3 class and sets the values for bank records, ad hub and public key, which are passed as parameters;
- function encrypt_records (records): this method is used to encrypt the input record data with the public key and returns the encrypted record;
- function intersect: this method finds an intersection between the bank records and the ad hub by using the encrypted values of the bank records and matching them with the corresponding values in the ad hub. If the values match, they are added to the result;
- function calculate_risk (epsilon): this method calculates the level of data privacy risk. It determines the size of the intersection between the bank records

and the ad hub and calculates the privacy budget based on the size of the bank records and the size of the ad hub. The result is used to calculate the data privacy risk level based on the epsilon parameter.

5 Results of Testing on Conditionally Real Data

Consider a typical data set (dataset), which often uses in organizations and contains information about the client (what, where, how and through what he purchased or viewed). The most common method for anonymization is using micro segments and masking.

Checking for dataset re-identification, based on information about purchases and viewing advertisements for useful and safe report at each processing step, using the k-anonymity model (When $K \geq 10$ - data is protected) During the conversion process, data converts several times - at the source level, through depersonalization (formation of micro-segments), and final depersonalization by micro-segments. There are key features of data storage organization:

- The object is used to solve analytical problems characterized by large data samples for machine learning and other advanced analytical tools. The data flow profile tends to be batch-oriented, the main users are internal analysis;
- The organization has a developed security system, the contextual risks calculated on the basis of calculators is 40% (0.4);
- Used depersonalization methods should allow protecting data in a recurrent manner (serial protection, repeated procedures) and make it possible to adjust the algorithm depending on attribute composition of the protected data and their volume;
- In fact, in exchange process three sets of data are processed and published: A (purchase data) - contains data tied to the phone number of an individual, B (advertising view data) - ad views data, also tied to the phone number, C (summary report) - formed by the merger of A and B, comparing purchases and ad views. The problem to be solved belongs to SMPC / PSI (Secure Multi-Part Calculation / Private Set Intersection) class;

In the data preparing process, they are anonymized: sets A - forms microsegments, then data is aggregated by purchases relative to the brand and product category; sets B - by ad views. Before sending formation of summary report, pointers to micro-segments are pseudonymised (reduced to abstract numbers). Further, combining reports A and B allows us to form set C, in which references microsegment aliases are also removed.

To create synthetic data (dummy data), uses an algorithmic data generator, tunes each attributes in the test set (7 sheets (A1, A2, A3, B1, B2, C1 and C2) with 500,000 data lines each). After choosing the degree of data blurring (degree of generalization), the results are:

- $k - anonymity = 1[94.39\%]$
- $k - anonymity = 2[5.61\%]$

After suppressing the rows $k - anonymity \leq 2$, the desired re-identification probability from the data is:

$$P_{data} = \frac{1}{1 * 2} = 0.5 \rightarrow P_{re-id} = 0.5 * 0.5 = 0.2 > 10\% \quad (5)$$

Thus, testing on synthetic data did not reach the safety threshold of 10%, which required refinement of the generalization parameters for testing on conditionally real data.

- The applied algorithm contains protective functions in relation to personal data;
- In terms of depersonalization procedures, other fields that are not depersonalized according to the proposed algorithm, such as the time of purchase, provide unique data accurate to the second;
- The overall k-anonymity metrics are low, mostly in the range of 1–4 and do not allow reducing the probability of re-identification below the threshold of 10%.

The depersonalization algorithm based on conditionally real data was largely based on the results obtained during the preliminary testing round. Key features of the approach used.The PSI3 algorithm works by dividing a large group into smaller subgroups and calculating the sum of each subgroup. It then adds Laplace noise to each of these subgroup sums to preserve differential privacy. It then computes the subgroup intersection by comparing the sums of the noisy subgroups. Finally, it adds more Laplace noise to the result of the intersection to further protect the privacy of the people in the subgroups.As a result of testing on conditionally real data, the following results were obtained:

- $k - anonymity = 8[98.239\%]$
- $k - anonymity = 16[1.727\%]$
- $k - anonymity = 24[0.035\%]$

After suppressing the rows $k - anonymity \leq 8$, the desired probability of re-identification from the data was:

$$P_{data} = \frac{1}{8} = 0.125 \rightarrow P_{re-id} = 0.125 * 0.4 = 0.05 \leq 10\% \quad (6)$$

Taking the data set and the k value as the k-anonymity metric of the input data, and first calculates the entropy of each attribute. If an attribute has non-zero entropy, it generalizes it using a masking method: replacing the last characters with wildcards (asterisks) until its entropy is less than or equal to the logarithmic base 2 of the value of k. If the attribute has high entropy (entropy greater than or equal to base 2 logarithm of k), it is added to the list of high entropy attributes that are potential quasi-identifiers Qi. (Fig. 2)

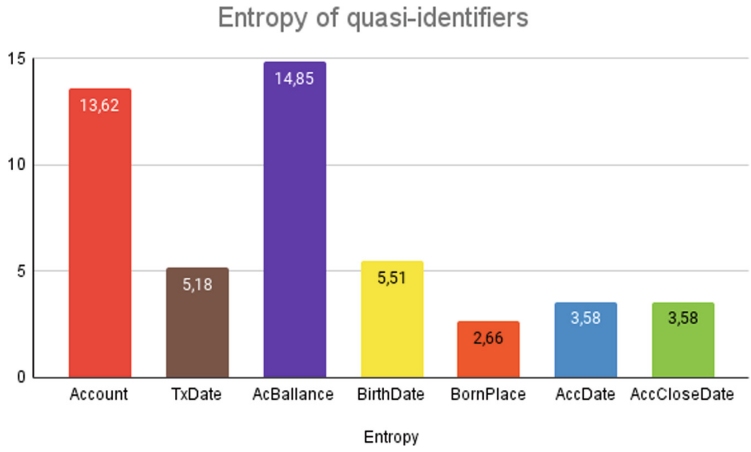


Fig. 2. Graphical interpretation of entropy on the original data

In a multi-threaded scenario using homomorphic encryption, this level of risk is reduced to a new level, namely

- $Intersectionsize = 49998$
- $Risklevel = 1.0002000400080017e - 05 = 0.0000100020$

$$P_{data} = 0.00001 \rightarrow P_{re-id} = 0.00001 * 0.4 = 0.000004 \leq 10\% \quad (7)$$

Thus, the use of additional degrees of generalization for conditionally real data made it possible to achieve the necessary threshold for the risk of re-identification.

6 Conclusions

The conducted testing allows us to formulate the following general conclusions on the use of depersonalization methods for large data structures:

1. Assessing the effectiveness of anonymization methods: It is important to evaluate the effectiveness of anonymization methods before publishing data. This can be done by measuring the degree of anonymization achieved using metrics, as well as checking whether it is possible to re-identify the individual from previously anonymized data.
2. Factoring the probability of reidentification: Explicit allocation of contextual risks reduces the probability of reidentification by protecting against multi-level attack vectors
3. Include in the depersonalization procedure all attributes of the set, including time, geolocation, and product categories. Because the sensitive attribute (representing the subject matter) is the purchase amount, which represents the aggregated value. This blurring of time and geo-location in a small range

will significantly improve risk indicators and will not affect the usefulness of the model.

4. Change the procedure for integrating banking data and ad viewing data to use advanced integration models:
   - Trusted intermediary model (for testing purposes). Limitation: Cannot be generalized.
   - Model of distributed learning (Federated Learning). Restriction: does not provide the ability to calculate metrics and quantitative risk assessments.
   - Confidential computing model combined with anonymization. Limitation: requires computational costs.
5. In the Confidential Computing Model The PSI3 Scheme with Homomorphic Encryption can be used to improve the security and privacy of communications over marketing attribution phone numbers. Compared to using microsegments on phone numbers, the PSI3 scheme provides greater accuracy and significantly better privacy. However, this is relatively slow and difficult to implement.
   - The PSI3 scheme allows financial institutions and marketing platforms to exchange customer phone numbers to determine the intersection of their respective sets without disclosing any additional information. This approach ensures the confidentiality and security of customer data, preventing any unauthorized access to confidential information.
   - Compared to using phone number microsegments, which generalize to the last two digits of the number, the PSI3 scheme provides greater accuracy in identifying common customers between two sets. The use of homomorphic encryption also ensures that the transmitted data remains confidential, further enhancing the privacy of customer information.
   - However, the PSI3 scheme can be relatively slow and complex to implement, requiring a significant amount of computational time and technical knowledge. Despite this, the accuracy and privacy advantages of this approach make it a promising solution for exchanging phone number data in marketing attribution

The article was prepared as part of the implementation of the program of strategic academic leadership “Priority-2030”, Strategic project No. 2 “Digital transformation of the university and industrial enterprises based on the Intelligent Cyber-Physical Platform”.

References

1. Bogdanov, A.V., et al.: Protection of personal data using anonymization. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12956, pp. 447–459. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87010-2_33
2. Bogdanov, A.V., et al.: Risk model of application of lifting methods. In: Proceedings of the 9th International Conference Distributed Computing and Grid Technologies in Science and Education. RWTH Aahen University, vol. 3041. pp. 369–374, December 2021. (CEUR Workshop Proceedings)

3. NISTIR 8062. An Introduction to Privacy Engineering and Risk Management in Federal Systems
4. ISO/IEC27701. Security techniques - Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management - Requirements and guidelines
5. De-identification Guidelines for Structured Data, IPC Ontario (2016)
6. General Data Protection Regulation (REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC)
7. Anonymization: managing data protection risk code of practice. Information Commissioner's Office, Wycliffe House, Water Lane, Wilmslow, Cheshire SK9 5AF. November 2012
8. ISO/IEC 20889:2018. Privacy enhancing data de-identification terminology and classification of techniques
9. Ito, K., Kogure, J., Shimoyama, T., Tsuda, H.: De-identification and encryption technologies to protect personal information. Fujitsu Sci. Tech. J. **52**(3), 28–36 (2016)
10. Shannon, C.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
11. Baez, J.C., Fritz, T., Leinster, T.: A characterization of entropy in terms of information loss. Entropy **13**, 1945–1957 (2011)
12. Wang, Z., Banawan, K., Ulukus, S.: Private set intersection: a multi-message symmetric private information retrieval perspective. IEEE Trans. Inf. Theor. **68**(3), 2001–2019 (2021)
13. Gómez-Sanz, Jorge J.., Pavón, Juan, Díaz-Carrasco, Áureo.: The PSI3 agent recommender system. In: Lovelle, Juan Manuel Cueva., Rodríguez, Bernardo Martín González., Gayo, Jose Emilio Labra., del Puerto Paule Ruiz, María, Aguilar, Luis Joyanes (eds.) ICWE 2003. LNCS, vol. 2722, pp. 30–39. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45068-8_5