Zoltán Alexin (2022): Entropy based approach to personal data. In: Proceedings of the International Conference on Privacy-friendly and Trustworthy Technology for Society – COST Action CA19121 - Network on Privacy-Aware Audio- and Video-Based Applications for Active and Assisted Living

Entropy based approach to personal data

Zoltán Alexin
University of Szeged, Department of Software Engineering, Hungary
alexin@inf.u-szeged.hu

Abstract

The forthcoming decade is envisioned being the era of Artificial Intelligence (AI). It seems that everything is at the disposal of the industry: the computing power, the large storage, and the Big Data. The latter is a key component of the AI systems, because the researchers' community considers that more training data result in more accurate AI software. Therefore, the industries demand larger and larger amount of genetic, biometric, health, geolocation, travel, and financial etc. data. The European Union and the member states try to keep the pace dictating the US and China and tend to satisfy the needs of the industries by laws, like the European Health Data Space (EHDS) Regulation, the AI Regulation. Applying these laws, the industries can get the much-needed data for themselves. What remains unsolved although, is the protection of individuals with regard to the automatic processing of personal data relating to them.

The requested data many times are personal data, at least once they were personal. Then underwent a de-identification procedure by which the natural identifiers were deleted. But it is not enough, because the data may still contain so-called quasi-identifiers by which an adversary can join two completely different datasets together and reveal the identity of the individuals whom the data relates to. When we talk about joining, it many times is understood in the general sense. That means, it can be executed based on proximity in time or geolocation, not only on the basis of identical values of some quasi-identifiers.

The author proposes a statistical method by which data protection experts can investigate datasets before handing over them to the industries. The method provides one single number, an entropy value, characteristic to the dataset that shows its vulnerability against re-identification attacks. The side effect of the method is that it provides distribution data over the quasi-identifiers, by which the analysts can identify the most and least vulnerable part of the population. The biggest risk is an adversary has a nationwide other dataset to attack with. To arm ourselves, we can model this kind of situation too by studying the entropy values and the distribution of quasi-identifiers at national level.

Introduction

Since the protection of personal data became a fundamental right in the European Union in 2009, when all member states undersigned the TFEU (Lisbon Treaty), it was always a question how we can decide that a particular dataset is personal at all. The law protects only personal data therefore such a judgement is crucial when a company, clinics or a public institution etc. want to share or publicize a dataset. In the case of health data, the rules of professional ethics also prohibit to reveal personal medical information before others.

The GDPR proposed a mean, by which the privacy rights of individuals can eventually be protected. It is the anonymization. The term assumes that the result of the process is an anonymous dataset. Since anonymity is questionable sometimes, therefore the *de-identification* is a more correct term. This latter means that the process intends to render the data anonymous, but it is not sure that this goal is achieved. Many cases were reported in the literature, for example in (Ohm, 2010) where the allegedly anonymous data later have been broken. This shows that the decision on anonymity must be especially cautious and be based on strong statistical evidence. The HIPAA law (US Federal Government, 2022) for example contains that a covered entity may determine that health information is not individually identifiable health information only if:

> § 164.514 b) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
>
> (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
>
> (ii) Documents the methods and results of the analysis that justify such determination

Such type of legal regulation is not known in the EU. According to the GDPR, the Data Protection Officer, who is generally a lawyer will decide on the anonymity of a dataset. The newly passed Hungarian Act XCI of 2021 on National Data Assets contains this approach too. In (Alexin, 2018) the author presented two court cases filed by himself on medical data protection. In one case the Supreme Court of Hungary after several appeals finally decided that the national Itemized Medical Dataset (IMD) does not contain personal data, although it contains a pseudonym (9-digit number), the birthdate, ZIP code, gender, dates of care, institutions, medicines related with a patient. Consequently, data subjects have no rights to access, to object, to be forgotten.

In (Higgins, 2021) the author argues that careful statistical analysis is essential before a research database is published. Datasets are analyzed from the point of view of k-anonymity and l-diversity. They quantified the risk for three scenarios:

- friendly researcher who might inadvertently reidentify an acquaintance
- a rogue researcher deliberately attempting reidentification using public information, and
- a rogue corporation with wide data access.

The future perspective is rather disappointing. Each day hackers stole a new personal database containing identification data for thousands if not millions of people. These databases sooner or later are being commercialized in the dark net. So, we must prepare for the case when any tiny identifier fragments in a de-identified dataset can be keys for re-identification. Some years ago, the data items in a dataset could be divided into two parts: quasi-identifiers and such type of data that are considered not suitable for re-identification. As hackers can have access to the original databases, we must accommodate to the fact that each data item will become quasi-identifier. In healthcare for example, the Electronic Health Record (EHR) systems holding all medical information from birth to death became a standard, therefore any tiny data item (heart rate, bilirubin concentration in urine, body weigh on a particular date) can be used for re-identification of patients either by an insider adversary (e. g. authority, researcher) or a hacker who stole original data records.

The anonymity therefore depends only on the amount of information about an individual stored in the de-identified dataset. The world population is 7.9 billion which corresponds to 33 bits of information, the population of Hungary is 10 million, it corresponds to 24 bits ($\log_2(population)$). See (Chang, 2019) which is a good attention-grabbing article on the topic. This amount of information is enough to identify somebody given the adversary can get such a personal database (a clue) that contains the complete or identical form of the quasi-identifiers exist in the de-identified dataset.

Computing the entropy of a dataset

In the probability theory a random variable X is a measurable function $X : \Omega \rightarrow E$ from a set of possible outcomes Ω to a measurable space E. In continuous case could be rather difficult but in the following the discrete case is applied. The Ω will denote a population. It can be a whole country, like Hungary, is a finite set. The random variable X randomly select an individual from the population and returns the quasi-identifiers of that individual as a tuple, like $(a, b, c, \dots)$, where $a \epsilon A, b \epsilon B, c \epsilon C$ etc. A could be the set of ZIP codes, where the individuals live, B could be the set of ages, C could be the set of genders. Such a way, $E = A \times B \times C \times \dots$. In our case $A, B, C, \dots$ all finite sets, therefore E also will become finite.

The probability of some $S \subseteq E$ is defined as

$$P(X \epsilon S) = P(\omega \epsilon \Omega : X(\omega) \in S) \tag{1}$$

It is assumed that every individual in the population is equally probable. Then $P(X \epsilon S)$ is proportional with the number of individuals whose quasi-identifiers are in S. Let the number of such individuals be k. S may contain one single element (tuple) from E. The values of P can be narrowed between *0* and *1* by dividing it with the number of elements in Ω. Let the number of elements in Ω *is* N. $|\Omega| = N$.

$$P(X \epsilon S) = \frac{|\{\omega \epsilon \Omega : X(\omega) \in S\}|}{|\Omega|} = \frac{k}{N} \tag{2}$$

This way $P(E) = 1, P(\emptyset) = 0$.

The probabilities of certain quasi-identifiers may differ. For example, the population in two ZIP code districts may differ substantially. If we select citizens fairly and randomly then the probability of choosing someone from a more populated ZIP code district is larger, as suggests formula (2).

Claude Shannon published his well-known formula in (Shannon, 1948) by which one can compute the information content of telecommunication messages.

$$E(Message) = -\sum_{i=1}^{n} P(\{x_i\}) \; log_2 \, P(\{x_i\}) \tag{3}$$

In his formula a $Message$ is composed of letters $x_1, x_2, \dots, x_n$. The probabilities of the letters are given in advance. With the above formula (3) he can determine the information content (entropy) of any message.

In this paper the author proposes a new application of the formula (3). The dataset being inspected is considered a $Message$. We assume that it contains quasi-identifiers of random citizens. The letters will be the quasi-identifiers (tuples). The

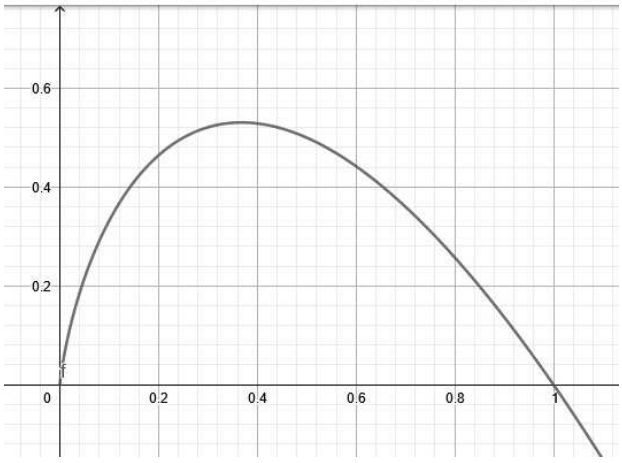
probabilities of the quasi-identifiers can be determined in advance. The best is to obtain data from national offices of statistics, or other official sources that are comprehensive and reliable. The application of the logarithm function can be reasoned as follows. If k individuals from the population share the same quasi-identifier (x_i) such a way that they cannot be distinguished from each other. In this case, the amount of information, the number of bits we can get to know is $\log_2(N/k)$. Because we can determine the group of k individuals having this quasi-identifier but cannot select one individual from the group. The $\log_2(N/k) = -\log_2(k/N) = -\log_2 P(\{x_i\})$.

One important note! In this paper a dataset $\mathcal{D}$ is considered a regular relational dataset, which means that one individual may occur at most once among the records. With this assumption the formula (3) provides a real and interpretable measure of unorderedness (entropy) over the quasi-identifiers of the population. Otherwise, the following considerations may not be true, for example, the amount of information accumulated in the dataset may increase indefinitely.

$$\mathcal{E}(E) = -\sum_{all\ e \epsilon E} P(\{e\})\ log_2 P(\{e\}) \tag{4}$$

On the other hand, when we have a snapshot of quasi-identifiers, we can compute the entropy for the whole population. During the computation a detailed insight to the distribution of the quasi-identifiers is also obtained. The distribution of the information gain can be characterized as well, for example we can tell the probability (number of individuals) of gaining at least n bit information about the individuals for a given n.

Figure I. The graph of the $f(x) = -x \log_2 x$ function



A plot of the function shows, that function $f(x)$ it can be computed only for positive real numbers. When $x = 0.0$ the right limit is 0.0, for $x = 0.25$ and $x = 0.5$ $f(x)$ will be 0.5 in both cases. The maximum $1/(e \ln 2)$ is reached when $x = 1/e$.

The author obtained a statistical research dataset from the population registry. The flowing computations all are demonstrated by this database. The research published here is considered a preliminary investigation.

What increases and decreases the entropy?

Lemma 1

Let N be a fixed integer number (e. g. the population taken), then the following inequality is held for all $0 < i, j, (i + j) < N$ integer numbers:

$$-\frac{i}{N}\log_2\frac{i}{N}-\frac{j}{N}\log_2\frac{j}{N}>-\frac{i+j}{N}\log_2\frac{i+j}{N} \tag{5}$$

Proof
If we substitute $a = i/N$ and $b = j/N$ then $0 < a, b, (a + b) < 1$ we get the following formula:

$$-a\log_2 a - b\log_2 b > -(a+b)\log_2(a+b) \tag{6}$$

which is equivalent

$$-\log_2 a^a b^b > -\log_2(a+b)^{(a+b)} \tag{7}$$

since a and b are positive numbers and less than 1, $1/a > 1/(a+b) > 1$

$$\log_2\left(\frac{1}{a}\right)^a\left(\frac{1}{b}\right)^b > \log_2\left(\frac{1}{a+b}\right)^a\left(\frac{1}{a+b}\right)^b = \log_2\left(\frac{1}{a+b}\right)^{(a+b)} \tag{8}$$

In fact, the inequality is held for all $0 < a, b$ values. If either of them, for example b is greater than 1, then the above formula can be modified as follows:

$$\log_2\left(\frac{1}{a}\right)^a\frac{1}{b^b} > \log_2\left(\frac{1}{a+b}\right)^a\frac{1}{(a+b)^b} = \log_2\left(\frac{1}{a+b}\right)^{(a+b)} \tag{9}$$

From the above lemma it follows for example, that

$$-\frac{1}{N}\log_2\frac{1}{N}-\frac{1}{N}\log_2\frac{1}{N}-\frac{1}{N}\log_2\frac{1}{N}-\frac{1}{N}\log_2\frac{1}{N}>-\frac{4}{N}\log_2\frac{4}{N} \tag{10}$$

$$-k\frac{1}{N}\log_2\frac{1}{N} > -\frac{k}{N}\log_2\frac{k}{N} \tag{11}$$

$$\log_2 N = -N\frac{1}{N}\log_2\frac{1}{N} \geq \sum_{i=1,2,\ldots,n}^{k_1+k_2+\cdots+k_n=N} -\frac{k_i}{N}\log_2\frac{k_i}{N} \tag{12}$$

The formula (12) says that the entropy of any discrete random variable must be less than or equal to $\log_2 N$ where N is the number of elements in Ω. The maximum value is reached if all individuals have different quasi-identifiers. In Hungary $\log_2(N)$ is ~23.254 bits. The formula (12) will result in 0, if all individuals belong to one single group of N indistinguishable individuals.

Lemma 2

The following inequality (13) is held for all $0 < k < l < N$ integer numbers. Having $k \cdot l$ individuals, such that they have l different quasi-identifiers, and for each quasi-identifier there exist exactly k individual who has this quasi-identifier. If we reverse the role of k and l, (k different quasi-identifier and l individual who has it) then the entropy will decrease. That means, the entropy can be decreased if we increase the number of indistinguishable individuals (from k to l).

$$-l\frac{k}{N}\log_2\frac{k}{N} > -k\frac{l}{N}\log_2\frac{l}{N} \tag{13}$$

$$-\frac{lk}{N}\log_2\frac{k}{N} > -\frac{kl}{N}\log_2\frac{l}{N} \tag{14}$$

The logarithm function is monotonic increasing, therefore the inequality (15) is held, because $k < l$. If we multiply both sides with the same negative coefficient $-kl/N$, then the direction of the inequality will reverse.

$$\log_2\frac{k}{N} < \log_2\frac{l}{N} \tag{15}$$

Corollary

$$-3\frac{3}{N}\log_2\frac{3}{N} > -\frac{4}{N}\log_2\frac{4}{N} - \frac{5}{N}\log_2\frac{5}{N} \tag{16}$$

We have 9 individuals in three groups such that we cannot distinguish them within a group. If we re-arrange them in two groups with 4 and 5 individuals such a way that they cannot be distinguished within a corresponding group, then the entropy will decrease. Lemma 2 is used twice for proving:

$$-4\ \frac{3}{N}\log_2\frac{3}{N} > -3\frac{4}{N}\log_2\frac{4}{N} \tag{17}$$

$$-12\ \frac{3}{N}\log_2\frac{3}{N} > -9\frac{4}{N}\log_2\frac{4}{N} \tag{18}$$

$$-12\ \frac{3}{N}\log_2\frac{3}{N} > -4\frac{4}{N}\log_2\frac{4}{N} - 5\frac{4}{N}\log_2\frac{4}{N} \tag{19}$$

$$-12\ \frac{3}{N}\log_2\frac{3}{N} > -4\frac{4}{N}\log_2\frac{4}{N} - 4\frac{5}{N}\log_2\frac{5}{N} \tag{20}$$

$$-3\ \frac{3}{N}\log_2\frac{3}{N} > -\frac{4}{N}\log_2\frac{4}{N}\ -\frac{5}{N}\log_2\frac{5}{N} \tag{21}$$

k-anonymity:

A dataset $\mathfrak{D}$ is considered k-anonymous for any natural number k, if for all records (representing a natural person) there exist at least $k-1$other record (individual) that they are indistinguishable from each other considering their quasi-identifiers.

Lemma 3

If a dataset $\mathfrak{D}$ is k-anonymous then its entropy $\mathfrak{E}(\mathfrak{D}) < -\log_2(k/N)$, where N is the number of individuals in the dataset.

$$-\log_2\frac{k}{N}\ \geq \sum_{i=1,2,\ldots,n}^{k_1+k_2+\cdots+k_n=N,\ \ k_i\geq k} -\frac{k_i}{N}\log_2\frac{k_i}{N} \tag{22}$$

Proof
It follows from the definition of k-anonymity and from Lemma 2. In the trivial case, when have exactly k individuals who share common quasi-identifiers and N is therefore divisible by k. Then we get the following formula:

$$\sum_{i=1,2,\ldots,N/k}^{k+k+\cdots+k=N} -\frac{k}{N}\log_2\frac{k}{N} = \frac{N}{k}\left(-\frac{k}{N}\log_2\frac{k}{N}\right) = -\log_2\frac{k}{N} \tag{23}$$

In general case, by applying Lemma 2 we can bring in new groups with $k+1, k+2, \ldots$ indistinguishable individuals, while doing this the entropy always decreases. A more elaborated formal proof is not available currently. See the Corollary also!

If the entropy of a dataset $\mathcal{D}$ is $\mathcal{E}(\mathcal{D})$, then we can compute an estimated k value, which is characteristic to the anonymity of the dataset.

$$-\log_2 \frac{k}{N} = \mathcal{E}(\mathcal{D}) \tag{24}$$

$$k = \frac{N}{2^{Entropy}} \tag{25}$$

From Lemma 3, it follows that the entropy of a 2-anonymous dataset is less or equal to $-\log_2(2/N) = \log_2(N/2) = \log_2(N) - 1$. An interesting question arose here: if we have a dataset $\mathcal{D}$ and its entropy falls between $\log_2(N) - 1 < \mathcal{E}(\mathcal{D}) < \log_2(N)$ then how many uniquely identifiable records (singletons) must exist in the database? The following equation system needs to be solved.

$$-\lambda_2 \frac{2}{N} \log_2 \frac{2}{N} - \lambda_1 \frac{1}{N} \log_2 \frac{1}{N} = \mathcal{E}(\mathcal{D}) \tag{26.1}$$

$$2\,\lambda_2 + \lambda_1 = N \tag{26.2}$$

By substituting λ_1/N by x the following equation is obtained:

$$(1-x)\,(\log_2(N) - 1) + x \log_2(N) = \mathcal{E}(\mathcal{D}) \tag{27}$$

$$\log_2(N) - 1 + x = \mathcal{E}(\mathcal{D}) \tag{28}$$

$$x = \mathcal{E}(\mathcal{D}) + 1 - \log_2(N) \tag{29}$$

$$\lambda_1 = (\mathcal{E}(\mathcal{D}) - (\log_2(N) - 1)) * N \tag{30}$$

It shows, that if the entropy is less than $\log_2(N) - 1$ then no singletons are guaranteed, but above this threshold the number of guaranteed singletons is increasing until it reaches N when the entropy becomes $\log_2(N)$. The number of singletons can be larger, if the dataset $\mathcal{D}$ contains, not only pairs, but couple of sets of three, four, five, … indistinguishable individuals.

$$n_{singletons} \geq (\mathcal{E}(\mathcal{D}) - (\log_2(N) - 1)) * N \tag{31}$$

The entropy computations with the Hungarian population registry dataset

The author obtained a research dataset from the Central Office for Administrative and Electronic Public Services (in Hungarian: Közigazgatási és Elektronikus Közszolgáltatások Központi Hivatala, KEKKH) which contained all Hungarian citizens' date of birth, ZIP code of his/her resident address and gender. Altogether 10 004 090 people were in it, $N = 10\,004\,090$. The dataset is a snapshot which have been taken on 31st December 2011 at midnight. Earlier, in (Alexin, 2014) the author investigated the dataset from the point of view of identifiability.

The **Bits** column is always defined as $\log_2(N/k)$. The value of **Entropy** is: $-k/N \log_2(k/N)$.

The entropy of the ZIP code quasi-identifier

Table I. Hungarian ZIP codes, population, and information content

ZIP code	Settlement	Population	Bits	Entropy
1011	Budapest I.	3286	11,5719	0,003800
1012	Budapest I.	4446	11,1357	0,004948
1013	Budapest I.	3404	11,5210	0,003920
…				
9982	Apátistvánfalva	589	14,0519	0,000827
9983	Szakonyfalu	769	13,6672	0,001050
9985	Felsőszölnök	589	14,0519	0,000827
Sum:		**10004090**		**10,303428**

The result of the computation shows that that the entropy of ZIP codes is 10,3 bits. It means that statistically, for a random citizen the expected amount of information in his/her ZIP code is 10.3 bits. It may be more or less since it is an average. It corresponds to 7916-anonymity using the formula (25). Therefore, the ZIP code alone does not mean any privacy risk in a database.

When we look at the Table II. closely, we see that although, the median is about 10 bits, the amount of information gained form a ZIP code ranges from 6 bits to 15 bits. The reason is that the population in a ZIP code district is rather imbalanced. There are sparsely and densely populated districts. The range is from 100 to 100 000. For 48 593 citizen the ZIP code means 15 bits (305-anonymity). It is an elevated but bearable risk.

Table II. The probability of gaining at least *n* bit information if someone's ZIP code is known

Bits	Population	Probability
15	48593	0,49%
14	302595	3,02%
13	966087	9,66%
12	2139699	21,39%
11	3476436	34,75%
10	5210515	52,08%
9	7369394	73,66%
8	8847670	88,44%
7	9716633	97,13%
6	10004090	100,00%

The entropy of the date of birth quasi-identifier

Table III. Hungarian birthdates' distribution, and their information content

Birthdate	Population	Bits	Entropy
1894.12.31.	1	23,25409	2,32446E-06
…			
1985.01.01.	306	14,996698	0,000458711
1985.01.02.	335	14,866069	0,000497810
1985.01.03.	365	14,742333	0,000537875
1985.01.04.	367	14,734450	0,000540533
1985.01.05.	331	14,883399	0,000492439
1985.01.06.	296	15,044633	0,000445139
…			
Sum:	**10004090**		**14,918582**

In this case the result of the computation shows that that the entropy is 14,918 bits. It corresponds to 323-anonymity using the formula (25). Therefore, the date of birth alone does not mean serious privacy risk in a database. The eldest citizen was born in 1894. according to the dataset. It can be seen, that among those people who was born at the beginning of 1985, usually ca. 300 were indistinguishable.

The distribution of the information gain is more balanced as shown in Table IV. It ranges from 13 to 17 bits. The number of indistinguishable citizens is decreasing year by year, but slowly. The number of births is quite stable and even. From the first line we can discover that in the case of 907 citizen the date of birth means unique identifiability (23 bits, 1-anonymity), for 1979 citizens the gain is 22 bits, 2-anonymity, for 4573 21 bits, 3- or 4-anonymity. The table helps us to recognize that we have a smaller, but vulnerable group of people who needs special attention.

The author computed the entropy of the year and month of birth quasi-identifier which resulted in 9.99 bits, 9837-anonymity. This way the risk can be substantially reduced, but these very old citizens remain still uniquely identifiable. Their data shall be suppressed before transferring.

Table IV. The probability of gaining at least *n* bit information if someone's date of birth is known

Bits	Population	Ratio
23	907	0,01%
22	1979	0,02%
21	4573	0,05%
20	10944	0,11%
19	15778	0,16%
18	38252	0,38%
17	117105	1,17%
16	378792	3,79%
15	3282589	32,81%
14	9994548	99,90%
13	10004090	100,00%

Cartesian products of certain quasi-identifiers

Table V. Hungarian $Birthdate \times ZIP\ code$ distribution and the information content

Birthdate x ZIP code	Population	Bits	Entropy
(1894.12.31., 3744)	1	23,254	2,324458e-6
…			
(1975.08.04., 9400)	4	21,254	8,498159e-6
(1975.08.04., 9407)	1	23,254	2,324458e-6
(1975.08.04., 9473)	1	23,254	2,324458e-6
(1975.08.04., 9523)	1	23,254	2,324458e-6
(1975.08.04., 9600)	1	23,254	2,324458e-6
(1975.08.04., 9700)	6	20,669	1,239640e-5
…			
Sum:	**10004090**		**22,79385**

The result of the computation shows dramatic changes when we examine the date of birth x ZIP code quasi-identifier. See Table V. The entropy became 22,7985 bits. It corresponds to 1.37-anonymity using the formula (25). This database poses substantial risk for re-identification. Must not be released or transferred. Using the formula (31) the ratio of singletons is greater the 54% of the population, in fact it was 6635838 individuals. This is clearly seen in Table VI.

Table VI. The probability of gaining at least *n* bit information if someone's date of birth and ZIP code are known

Bits	Population	Ratio
23	6635838	66,33%
22	8629982	86,26%
21	9692881	96,89%
20	9996707	99,93%
19	10004090	100,00%

Table VII. Hungarian $birthdate \times ZIP\ code \times Gender$ distribution and the information content (M – male, F – female)

Birthdate x ZIP x gender	Population	Bits	Entropy
(1894.12.31., 3744, M)	1	23,254	2,324458e-6
…			
(1954.04.14., 6000, M)	1	23,254	2,324458e-6
(1954.04.14., 6041, M)	1	23,254	2,324458e-6
(1954.04.14., 6066, M)	2	22,254	4,448998e-6
(1954.04.14., 6070, F)	1	23,254	2,324458e-6
(1954.04.14., 6097, F)	1	23,254	2,324458e-6
(1954.04.14., 6097, M)	1	23,254	2,324458e-6
…			
Sum:	**10004090**		**22,992721**

The last computation shows even dramatic changes when we examine the date of birth x ZIP code x gender quasi-identifier. See Table VII. The entropy became 22,9927 bits. It corresponds to 1.19-anonymity using the formula (25). This database poses substantial risk for re-identification. Using the formula (31) the ratio of singletons is greater the 74% of the population, in fact it was 7845850 individuals. This is seen in Table VIII.

Table VIII. The probability of gaining at least *n* bit information if someone's date of birth and ZIP code and gender are known

Bits	Population	Ratio
23	7845850	78,43%
22	9403904	94,00%
21	9942428	99,38%
20	10003959	99,99%
19	10004090	100,00%

Summary

This article presented preliminary research on the entropy of certain quasi-identifiers, and combination of quasi-identifiers based on reliable statistical data. The computations can be repeated by other national databases with other quasi-identifiers. The presented approach could be an ultimate decision-support tool for data guardians before they decide on the transfer of a dataset to third parties.

Acknowledgments

The author wishes to thank for the support of the COST Action CA19121 Network on Privacy-Aware Audio- and Video-Based Applications for Active and Assisted Living *"GoodBrother"* project.

The author would like to express his thanks to the Central Office for Administrative and Electronic Public Services (in Hungarian: Közigazgatási és Elektronikus Közszolgáltatások Központi Hivatala, KEKKH) for the research dataset extracted from the national population registry.

References

Alexin, Z. (2014): Does fair anonymization exist? International Review of Law, Computers and Technology, Taylor & Francis Publishing, Vol. 28 No. 1: pp. 21-44, DOI: 10.1080/13600869.2013.869909

Alexin, Z. (2018): Court cases relating to medical data protection, Interdiszciplináris Magyar Egészségügy, Larix Kiadó Kft., in Hungarian, Vol.: XVII. No.:3, pp. 57-62

Chang, Kenny (2019): Personal Data and 33 bits of Entropy, SynchLab webpage: https://synch.law/personal-data-and-33-bits-of-entropy/

Ohm, P. (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization UCLA Law Review, University of Colorado, Law Legal Studies Research Paper Vol. 57, No. 9-12, p. 1701, Available at SSRN: https://ssrn.com/abstract=1450006

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal. 27 (3): 379–423, 623–656. doi:10.1002/j.1538-7305.1948.tb01338.x.

Sweeney, L. (2000) Simple demographics often identify people uniquely, Data Privacy Working Paper 3. Pittsburgh Carnegie Mellon University, http://dataprivacylab.org/projects/identifiability/paper1.pdf

Higgins T. L. (2021): Reidentification of Protected Health Information: Can the Risk Be Quantified? Crit Care Med. 2021 Jun 1;49(6):1003-1006. doi: 10.1097/CCM.0000000000004931. PMID: 34011836.

US Federal Government (2022). 45 CFR 164.514 Other requirements relating to uses and disclosures of protected health information (HIPAA Law, The privacy rule). https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-E