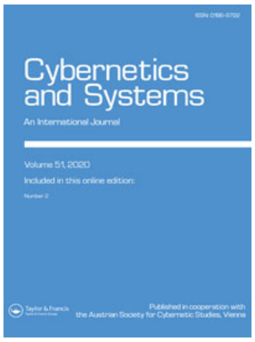

ISSN 0196-9722

Cybernetics and Systems
An International Journal

Volume 51, 2020
Included in this online edition:
Number 2

Taylor & Francis
Published in cooperation with the Austrian Society for Cybernetic Studies, Vienna

Taylor & Francis
Taylor & Francis Group

Cybernetics and Systems

An International Journal

ISSN: 0196-9722 (Print) 1087-6553 (Online) Journal homepage: https://www.tandfonline.com/loi/ucbs20

The Use of Context-Free Probabilistic Grammar to Anonymise Statistical Data

Zygmunt Mazur & Janusz Pec

To cite this article: Zygmunt Mazur & Janusz Pec (2020) The Use of Context-Free Probabilistic Grammar to Anonymise Statistical Data, Cybernetics and Systems, 51:2, 177-191, DOI: 10.1080/01969722.2019.1705551

To link to this article: https://doi.org/10.1080/01969722.2019.1705551

Published online: 10 Jan 2020.

CYBERNETICS AND SYSTEMS: AN INTERNATIONAL JOURNAL
2020, VOL. 51, NO. 2, 177–191
https://doi.org/10.1080/01969722.2019.1705551

The Use of Context-Free Probabilistic Grammar to Anonymise Statistical Data

Zygmunt Mazur[a] and Janusz Pec[b]

[a]Faculty of Computer Science and Management, Wroclaw University of Science and Technology, Wrocław, Poland; [b]Central Statistical Office, Warsaw, Poland

ABSTRACT

In the following article, a proprietary method of anonymisation of identifiable statistical data using context-free probabilistic grammar is proposed. The advantage of this method is that it is simple and thanks to this, the identifier is easy to retrieve after masking the identifiable data, e.g. when it is necessary to modify or update the micro-data. This can be done using public-key cryptography, i.e. encrypting some probabilistic context-free grammar with this method. In the case of public statistics, there is often a need to use an anonymised source value, for example when economic operators' reports are verified by statistical officers. With appropriate information generated by context-free grammar, the verifier can easily identify an economic operator or a natural person. The idea of the anonymising algorithm used in the proposed method is presented by means of an example. According to the authors, the combination of the proposed method with asymmetric encryption of the definition of context-free grammar using public key infrastructure, makes it probable that its resistance to attacks will be quite high. This is because statistical methods that are used in the analysis of natural languages are not susceptible to attacks.

KEYWORDS

Anonymisation; argument tree; asymmetric encryption; context-free grammar; electronic notary public; identifiable statistical micro-data; medical personal data; pseudonymisation; security policy

Introduction

In statistical surveys, the protection of identifiable micro-data – including sensitive data – is one of the most important issues related to the management of information security in official statistics. Identifiable information relating to an object that is a natural person or a legal entity can be stored in various forms in databases or social networks. It should be noted that the anonymisation methods used in the field of personal data can be successfully applied to protect individual statistical data. The protection of such data, in Polish and also European Union law, is strengthened by so-called statistical confidentiality, if personal data are collected for statistical purposes. In the case of social

CONTACT Zygmunt Mazur zygmunt.mazur@pwr.edu.pl Faculty of Computer Science and Management, Wroclaw University of Science and Technology, Wrocław, Poland.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ucbs.
This article has been republished with minor changes. These changes do not impact the academic content of the article.

networks, different methods of data anonymisation are used. Different procedures for anonymisation of micro-data are used in public statistics – there is no uniform established and formalized procedure, but there are general guidelines defined by Eurostat, the European Union's statistical agency. In addition, there is no uniform and precise definition of the concept of "anonymisation of data" that is acceptable to all stakeholders. Polish law in Article 3(1) of the Act of 16 September 2011 on the exchange of information with law enforcement authorities of the Member States of the European Union, Journal of Laws 2011 No. 230, item 1371) (Act 2011) defines anonymisation as the transformation of personal data in such a way that it is impossible or disproportionate in terms of costs, time or activities to attribute specific information to an identified or identifiable natural person. A narrower term for anonymisation can be found in the EU document (Guidance Note: Guidance on Anonymisation and Pseudonymisation 2019).

Anonymisation of data is not a trivial process, as some authors state, e.g. in paper (Gruschka et al. 2018), especially if the requirements of the EU regulations are to be met (Mostert et al. 2016). Problems related to the practical reduction of data privacy risk are described in paper (Whitelegg 2018), and methodological guidelines on data privacy in the cloud environment are presented in the EU document (Guidelines on the use of cloud computing services by the European institutions and bodies 2018). A partial discussion of privacy risks regarding "Big Data" can be found in paper (MSI-NET 2016).

Many public statistical services use tools embedded in commercial software to anonymise data – these are the relevant software modules contained in commercial databases. Some people use methods they have developed. Because the quality of methods used is different, it seems advisable to develop and propose a new scheme of anonymisation procedure. This issue is the subject of this article.

Due to the necessary high quality of data anonymisation methods used in statistics, the following questions need to be answered:

(1) What methods of anonymisation are currently used in the organization concerned (in the company, office, foundation, etc.) when performing planned tasks?
(2) Do these methods have any security certificates and if so, which ones?
(3) Are these methods documented in the form of precisely described algorithms, e.g. in the form of pseudo-code, graphic code (as block diagrams) or in the source code in selected programming language in which the algorithm is implemented.

When using anonymisation methods based on pseudo-random number generators, it is necessary to provide the following information:

- Generator type.
- Theoretical period of repeatability of number generation.
- Other parameters characterizing the operation of the generator.
- Tests carried out for a given generator (e.g. according to the American standard FIPS 140-2 (FIPS 1999) there are four basic tests constituting the test minimum: monobite, poker, series, long series).
- Additional tests, e.g. spectral, linear complexity, sequential (Wald 1945), entropic.
- Frequency of updating assumptions for generator tests.

There are many methods of anonymisation. A review of selected methods (mainly used in medicine) is included in the works (Liber 2014a, 2014b; Mostert et al. 2016).

Pseudonymisation is a method related to anonymisation. Some authors classify pseudonymisation as one of the forms of anonymisation, as it was assumed in paper (Borucki 2009). However, due to the differences in the techniques used, EU experts separate the two methods (Opinion 05/2014 on Anonymisation Techniques 2014). It should be noted that so far, there is no formal definition of pseudonymisation in EU documents. The EU view on the problems of anonymisation is well reflected in the "Opinion 05/2014 on Anonymisation Techniques" (2014). When discussing this issue, one should also mention the use of anonymisation techniques in social networks. Due to the specificity of social networks, which are more complex than database systems, the techniques used there differ significantly from those used in relational database systems (Tripathy et al. 2012).

In medicine, an important issue in the protection of micro-data is the different structure of patient data. In the works (Borucki 2009; Liber 2014a, 2014b), medical data of patients are treated as medical personal data. Apart from text data, image data (tomography, magnetic resonance imaging, ECG recordings, positronic tomographic emission – PET) are widely used here. Paper (Borucki 2009) describes one of the simplest and most effective models of anonymisation methods – the data separation model. It is based on the division of data into two groups – sensitive data are separated from other data. The separation process should be carried out precisely so that, firstly, the identification data does not include any redundant information, and secondly, so that after separation the remaining data is useful for further processing, e.g. statistical processing. This method assumes the introduction of the definition of a notary – physical or electronic supervisor of the database, i.e. a dictionary that should be particularly well-protected. For more complex cases, including those belonging to the so-called critical information structure of a particular medical organization (e.g. hospital), multi-layered databases of connections are proposed. This certainly includes cases where an attacking information system may use data from external databases in relation to data within the organization.

Review of Commonly Used Anonymisation Techniques

The Data Protection Working Party, established by the European Commission Directive –Article 29 of Directive 95/46/EC (European Parliament and Council of the European Union 2016), as an advisory body to the Commission in the field of data protection and implementation of privacy protection rules, has identified three basic aspects of the assessment of anonymisation techniques, namely:

- Singling out – whether it is possible to separate all or some of the data identifying a natural person in a data filing system.
- Linkability – links between records (in one or more databases) – whether an attacker is able to assign two records to the same group.
- Inference – whether other data can be deduced from certain data (attributes) with high probability.

Each of these aspects of the anonymisation technique is taken into account when evaluating individual methods. Use of each of these methods involves a certain risk because each of them is exposed to human errors which are common in practice. A brief discussion of these errors can be found in the EU document (Opinion 05/2014 on Anonymisation Techniques 2014).

The European Commission's working group has identified four main methods for anonymising data:

- Randomization
- Generalization
- Pseudonymisation
- Anatomization – of data separation.

Each of the above-mentioned methods has specific techniques, namely:

- A randomization method can be implemented through:
  - ○ Noise addition technique
  - ○ Permutation technique
  - ○ Technique for personalized privacy
- The generalization method can be implemented through:
  - ○ Aggregation technique,
  - ○ K-anonymity technique
  - ○ L-diversification technique

Table 1. Comparison of selected anonymisation techniques with regard to the possible three types of risk.

Name of the technique	Risk related to the awarding of distinctions	Risk associated with links between database records	Risk associated with the application
Pseudonymisation excluding hashing/tokenisation technology	Yes	Yes	Yes
Adding noise	Yes	Under certain conditions	Under certain conditions
Permutation	Yes	Yes	Under certain conditions
Aggregation/K-anonymisation	No	Yes	Yes
L-diversification	No	Yes	Under certain conditions
Personalized privacy	Under certain conditions	Under certain conditions	Under certain conditions
Hashing/Tokenisation	Yes	Yes	Under certain conditions

T-proximity technique

- The pseudonymisation method may be implemented through:
  - Dictionary method
  - Hashing
  - Tokenisation
  - Symmetric encryption technology
  - Asymmetric encryption technology (with public and private keys)
- Anatomization (data separation)
  - The data separation model consists of dividing data into two parts – separating sensitive data from other data. The separation process must be carried out in such a way that the isolated identification data does not contain any redundant information but only identification information. The relationship between the two types of data is defined by pseudonyms. The idea is, therefore, the same as for pseudonymisation, but with a dictionary.

Table 1 shows a comparison of selected anonymisation techniques taking three basic aspects of their evaluation into account.

A more detailed discussion of the techniques mentioned here can be found in papers (Borucki 2009; Liber 2014a, 2014b), as well as in the works of the authors of the individual methods. Discussion of some anonymisation methods contained in commercial packages can be found in paper (Nabywaniec 2019), and the problem of finding a balance (compromise) between the usefulness of information after anonymisation and the level of its privacy is discussed in papers (Danilowicz and Nguyen 2000; Yu 2016). These papers enable the evaluation of various methods of anonymisation. Many articles, e.g. Liber (2014a, 2014b), devoted to methods of protection of micro-data concern the protection of patients' medical data as particularly sensitive and attractive to attackers, and thus are exposed to external attacks. Issues of protection of medical data are also of particular interest for health statistics.

Proposal for a Method of Anonymisation of Micro-Data Using Probabilistic, Context-Free Grammar

In this section, we will present a proprietary method of anonymising individual data using the properties of context-free grammar. Here are two basic definitions that will be used when discussing this method.

(1) Context-free grammar is called formal grammar of type 2 according to Chomsky's hierarchy, i.e. ordered four (T, N, P, S), where:
- T is a finite collection of terminal symbols,
- N is a finished collection of symbols of nonterminal Ni,
- P is a finite set of transcription rules L → R, L $\in$ N, R $\in$ (T $\cup\mathbf{N})^*$,
- S $\in$ N is a distinguished initial symbol.

(2) Probabilistic, context-free (Probabilistic Context-Free Grammar – PCFG) is a context-free grammar that includes the probabilities of its production rules and is denoted by the symbol $\mathbf{G}_{\mathbf{B}}^{\mathbf{P}}$. Production probabilities are assigned by observing that the sum of probabilities of rules with the same predecessor is 1.

By denoting any set of terminal or nonterminal symbols by μj, the above condition on the rule probability for any nonterminal symbol Ni can be expressed as follows:

$$\sum_{j} \boldsymbol{P}\ (\boldsymbol{N}_i\ \rightarrow \mu_j) = 1$$

And P(Ni → $\boldsymbol{\mu}_j$) should be treated as a conditional probability P(Ni → $\boldsymbol{\mu}_j|\boldsymbol{N}_i$), while i, j are natural numbers, i is an index of nonterminal symbols, j is an index of sequences of terminal or nonterminal symbols.

The role of probability in studies of formal grammar, including context-free grammar, boils down to the study of statistical properties of natural and artificial languages. In this way, one can model and study the probabilistic properties of languages generated by grammar, for example, determine what the probability is of obtaining a given final sequence of characters from the initial symbol. For this reason, a function is introduced in which value on each rule should reflect the probability that the rule will be applied to the output. In contextual grammars, the probabilistic approach to rules reflects subtle differences in the use of language phrases depending on the circumstances and the current situation, i.e. the context.

Let us consider an example illustrating a method of data anonymisation inspired by paper (Liber 2014b). Table 2 presents the sensitive data of patients.

Let's define some probabilistic context-free grammar $\mathbf{G}_{\mathbf{B}}^{\mathbf{P}}$, where:

Table 2. Example of sensitive personal data.

No.	Forename	Surname	Sex	City	Profession	Disease
1	John	Biden	M	London	Engineer	AIDS
2	Alex	Brown	M	London	Engineer	AIDS
3	Jack	Wilson	M	London	Engineer	Influenza
4	Oscar	Taylor	M	London	Engineer	AIDS
5	Thomas	Byrne	M	New York	Painter	Cancer
6	Peter	Anderson	M	New York	Painter	Influenza
7	Emily	Aster	F	Paris	Musician	Cancer
8	Alice	Morton	F	Paris	Dancer	AIDS
9	George	O'Connor	M	Paris	Musician	AIDS

- T – a set of terminal symbols, i.e. a set of characters present in the data in Table 2, i.e. in columns ***Forename, Surname, Sex, City, Profession, Disease.***

T = {a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, t, u, w, y, z, ‘} ∪ {ε}, where ε – empty symbol.

For the sake of simplicity, we assume that the letters are not case sensitive.

- **N = {Sex, City, Profession, Disease}**
- S – initial symbol, S = ***ForenameSurname***.

For the purpose of the example from Table 2, we take the first name of the Surname as the initial symbol (person identifier), as this does not lead to ambiguity in this case. However, in the case of ambiguity of the identifier, i.e. if there are two people with the same name and surname, other attributes that uniquely identify persons should be taken into account, e.g. PESEL number, which identifies each Polish citizen. In other countries, their national databases can be used, which unambiguously identify a citizen of a given country or persons legally staying abroad, e.g. by SSN (Social Security Number) in the USA. Of course, this applies to countries that have or will create such databases. We should also take into account that errors may appear in such databases. However, these databases are gradually being cleared of errors. If there are no such databases of citizens, and when we find ambiguity when adding data of people, we can extend the identifier to other fields, e.g. "Date of birth", "Place of birth", "Mother's maiden name" or other additional data, gradually removing the ambiguity from the database. Such practices are used, for example, when authenticating customers in a banking system.

P – a set of rewriting rules (productions), in which the "|" sign indicates an alternative (word "or"), and the normal fractions in the top right-hand corner of each product indicate the likelihood of their occurrence and their sum is 1.

Set of rules **P**:

1. S = ***ForenameSurname*** → ***SexCityDiseaseCityProfession***^{1}
2. ***Sex*** → k$^{1/8}$ | m$^{1/8}$ | k***Sex***$^{1/8}$ | m***Sex***$^{1/8}$ | ε$^{1/4}$ | ***CityProfession***$^{1/8}$ | ***DiseaseCity***$^{1/8}$
3. ***Disease*** → a$^{1/14}$ | d$^{1/14}$ | g$^{1/14}$ | i$^{1/14}$ | n$^{1/14}$ | o$^{1/14}$ | ‘$^{1/14}$ | p$^{1/14}$ | r$^{1/14}$ | s$^{1/14}$ | t$^{1/14}$ | w$^{1/14}$ | y$^{1/14}$ | ε$^{1/14}$
4. ***Profession*** → a$^{1/13}$ | e$^{1/13}$ | i$^{1/13}$ | l$^{1/13}$ | m$^{1/13}$ | n$^{1/13}$ | r$^{1/13}$ | t$^{1/13}$ | u$^{1/13}$ | y$^{1/13}$ | z$^{1/13}$ | ‘$^{1/13}$ | ε$^{1/13}$
5. ***City*** → a$^{1/28}$ | b$^{1/28}$ | e$^{1/28}$ | g$^{1/28}$ | i$^{1/28}$ | k$^{1/28}$ | l$^{1/28}$ | o$^{1/28}$ | ‘$^{1/28}$ | p$^{1/28}$ | r$^{1/28}$ | w$^{1/28}$ | z$^{1/28}$ | ε$^{1/28}$ | a***Sex***$^{1/28}$ | b***Sex***$^{1/28}$ | e***Sex***$^{1/28}$ | g***Sex***$^{1/28}$ | i***Sex***$^{1/28}$ | k***Sex***$^{1/28}$ | l***Sex***$^{1/28}$ | o***Sex***$^{1/28}$ | ‘***Sex***$^{1/28}$ | p***Sex***$^{1/28}$ | r***Sex***$^{1/28}$ | w***Sex***$^{1/28}$ | z***Sex***$^{1/28}$ | ε***Sex***$^{1/28}$

Having defined above a probabilistic, context-free grammar we can, for example, anonymise item 9 from the table (S = George O'Connor), using the sequence of production of the form c1.c2.c3…, where c1 means the position of production in the set P, and the remaining digits mean the sequence of production in a given item (if there is more than one). For example, we may obtain the following sequence of production as below:

S = George O'Connor → ***SexCityDiseaseCityProfession***
– (2.6) → ***CityProfessionCityDiseaseCityProfession***
– (5.21) → l***SexProfessionCityDiseaseCityProfession***
– (2.5) → l***ProfessionCityDiseaseCityProfession***
– (4.1) → la**CityDiseaseCityProfession**
– (5.1) → laa**DiseaseCityProfession**
– (3.5) → laan***CityProfession***
– (5.15) → laana ***SexProfession***
– (2.5) → laana***Profession***
– (4.5) → laanam

In this way, we assigned to the initial symbol S = George O'Connor the pseudonym "laanam". This represents a sequence of digits in the form of c1.c2c3 separated by spaces or commas. The returned sequence of digits for the above example is as follows (1, 2.6, 5.21, 2.5, 4.1, 5.1, 3.5, 5.15, 2.5, 4.5). The top-down left-hand argument tree of the pseudonym "laanam" for S = George O'Connor is as shown in Figure 1, where the numbers of the production rule used are given next to the arrow symbol in brackets.

Similarly, for the remaining data from Table 2, we can obtain further pseudonyms according to the proposed method, using the defined context-free probabilistic grammar and the top-down strategy of the

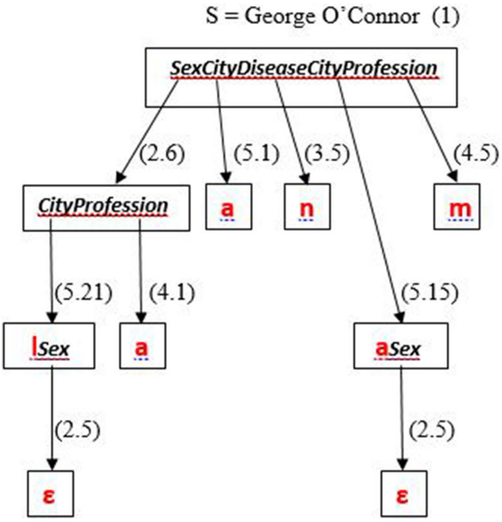

Figure 1. Left-hand tree of the pseudonym "laanam".

left-hand argument. The full list of their derivation in the form of sequences of digits describing the sequence of used production rules is as follows:

1. S = John Biden → (1, 2.3, 2.5, 5.15, 2.2, 3.1, 5.14, 4.11) → kamaz
2. S = Alex Brown → (1, 2.7, 3.14, 5.2, 5.20, 2.3, 2.5, 3.7, 4.8) → bkk't
3. S = Jack Wilson → (1, 2.7, 3.12, 5.8, 5.13, 5.14, 5.15, 2.1, 4.13) → wozak
4. S = Oscar Taylor → (1, 2.7, 3.5, 5.3, 5.27, 2.5, 3.3, 5.20, 2.4, 2.1, 4.13) → nezgkmk
5. S = Thomas Byrne → (1, 2.4, 2.4, 2.5, 5.2, 3.13, 5.10, 4.1) → mmbypa
6. S = Peter Anderson → (1, 2.5, 5.15, 2.4, 2.5, 3.12, 5.21, 2.4, 2.5, 4.13) → amwlm
7. S = Emily Aster → (1, 2.5, 5.19, 2.4, 2.5, 3.1, 5.14, 4.7) → imar
8. S = Alice Morton → (1, 2.3, 2.3, 2.5, 5.27, 2.5, 3.7, 5.17, 2.4, 2.5, 4.13) → kkz'em

Table 3. Modified data from Table 2 with pseudonyms.

No.	Pseudonym	Sex	City	Profession	Disease
1	kamaz	M	London	Engineer	AIDS
2	bkk't	M	London	Engineer	AIDS
3	wozak	M	London	Engineer	Influenza
4	nezgkmk	M	London	Engineer	AIDS
5	mmbypa	M	New York	Painter	Cancer
6	amwlm	M	New York	Painter	Influenza
7	imar	F	Paris	Musician	Cancer
8	kkz'em	F	Paris	Dancer	AIDS
9	laanam	M	Paris	Musician	AIDS

Table 4. Probability of occurrence for the 32 letters of the Polish language.

A	Ą	B	C	Ć	D	E	Ę
0.086	0.011	0.012	0.039	0.005	0.033	0.079	0.010
F	G	H	I	J	K	L	Ł
0.050	0.013	0.010	0.079	0.024	0.030	0.021	0.014
M	N	Ń	O	Ó	P	R	S
0.028	0.060	0.002	0.071	0.008	0.028	0.044	0.041
Ś	T	U	W	Y	Z	Ź	Ż
0.007	0.040	0.022	0.045	0.040	0.054	0.0006	0.007

After performing the above-defined operations, we get the following modified version of Table 2 after anonymisation (Table 3).

In the example, for the sake of simplicity of calculation, equal probability is assumed for production except for production with number 2.5, where the probability of producing an empty symbol ε is ¼. Any other value from the range of [0.1] can of course be used, e.g. the frequency of occurrence of letters in the natural language can be used – for example, in Polish (including diacritical characters) or in English, provided that the sum of the probabilities is 1. Probability values of occurrence of Polish letters in various texts are given in Table 4, and for English letters in Table 5 (Grajek and Gralewski 2009; Simon 2001). Of course, this will not reduce the security of the pseudonym, as in general, a pseudonym is a string of characters without semantics – unlike what is accepted in natural languages.

Information about the frequency of occurrence of letters in a given national language is often used in cryptography to break ciphers using statistical methods. First, these methods were used for cryptanalysis of monoalphabetic ciphers. For this purpose, the frequency of occurrence of individual letters of the code alphabet in the cipher was calculated and the values determined were compared with the frequency of occurrence of characters in the natural language. As we know, the analysis of the frequency of occurrence of letters depends on several factors, including the style of the individual author, the time period when the text was written, the subject matter it describes, the local dialect and other conditions. Therefore, we may get slightly different values for the frequency of occurrence of individual signs. However, these fluctuations decrease with the increase in the volume of samples of the analyzed texts – then they stabilize to a certain extent and for practical purposes they are precise enough.

Let us note that the form of production is arbitrary, it is subject only to limitations resulting from the definition of probabilistic, context-free grammar. In the presented example, if there is no "explicit" production of a given letter, then we assume that the probability of its generation is 0 (e.g. for the letter c). It should also be noted that each individual

Table 5. Likelihood of occurrence of the 26 letters of the English language.

A	B	C	D	E	F	G	H	I	J
0.082	0.015	0.028	0.043	0.127	0.022	0.020	0.061	0.070	0.020
K	L	M	N	O	P	Q	R	S	T
0.080	0.040	0.024	0.067	0.075	0.019	0.001	0.060	0.063	0.091
U	V	W	X	Y	Z				
0.028	0.010	0.024	0.020	0.200	0.001				

production is independent of each other, which results, for example, in the fact that in our example the probability of generating the letter “a” is different: for production number 3.1 it is equal to 1/14, for production number 4.1 it is 1/13 and for production number 5.1 it is equal to 1/28.

Discussion and Conclusions

Summarizing the considerations made in the previous sections and analyzing the specific properties of context-free and probabilistic grammar, one should pay attention to the computational complexity of the above method of anonymisation and its advantages and disadvantages, including susceptibility to attacks.

Note that the computational complexity of the method is of the order O(k2), because if there are n identifiers in the table and respectively for each identifier idj (where j = 1, 2, …, …, n) we used mj production rules, then

$$\sum_{1}^{n} \boldsymbol{m_j} \leq \mathrm{k}\cdot\mathrm{k}, \text{ where } \mathrm{k} = \max\{\mathrm{n}, \mathrm{m}_1, \mathrm{m}_2, \ldots, \mathrm{m}_\mathrm{j}, \ldots, \mathrm{m}_\mathrm{n}\}.$$

The main advantage of this method is the low cost of obtaining a pseudonym. In addition, an advantage is the obvious fact that, unlike deidentification, it makes it possible to pair different data related to a given identifier (e.g. a legal or natural person), while maintaining the condition of anonymity. Having additional information in the form of a sequence of digits separated by a space or a comma representing a sequence of production rules, we can create a reversible pseudonymisation, which guarantees a return from the pseudonymised data to a given identifier being e.g. a natural person – a customer, a patient, a sponsor, etc. We also create a reversible pseudonymisation. If we use asymmetric encryption. If we use asymmetric encryption, e.g. an AES system with a pair of keys(public, private) of the appropriate length for recording a given grammar, even if a hacker takes over a sequence of digits during data transmission – with the condition that the private key is securely stored – it is very difficult or even impossible to compromise the pseudonymised data. Moreover, the use of statistical methods is useful in the analysis of natural languages by analyzing the frequency of occurrence of letters in a given language is of little use because pseudonyms, in general, do not have to have any meaning – they are characterized by a lack of semantics. One of the additional advantages of this method is the opportunity to calculate the probability of the pseudonym for a given initial symbol S. For example, for the pseudonym tree "laanam" for S = George O'Connor and treating the individual production rules as independent events, we get the following probability:

$$\begin{aligned}&P(\text{argumentof}''\text{laanam}''\text{lettersequence}) = 1 \cdot 1/8 \cdot 1/28 \cdot 1/14 \cdot 1/8\cdot\\ &1/13 \cdot 1/28 \cdot 1/4 \cdot 1/4 =\\ &= \frac{1}{146112512} \leq \frac{1}{100000000} = 0.000000001.\end{aligned}$$

As far as the weaknesses of the presented method are concerned, they certainly include lack of clarity of context-free grammar. For context-free grammars, this problem is generally unresolvable. If the grammar is unambiguous, the derivation trees unambiguously define the structure (syntax) of words from the language. Of course, the final sequences of characters (i.e. words – replacement term) may have many leads, but they differ only in the order of applied production rules, and not in the structure of the word lead. The number of different derivations of a given word constitutes a certain imperfection of a given grammar in the aspect of the presented anonymisation method, because in the case of our method it increases the chances of an external attack on the pseudonym being discredited. This fact in grammar is often called grammar redundancy. It should be noted, however, that in other applications, ambiguity of context-free grammar has advantages, e.g. in the process of compiler construction.

Now, using the following notation, we will carry out the following simple reasoning. We denote:

- w_1, w_2, …, w_i, i = 1, 2, …, n, where w_i – the i-th argument of the word v – a pseudonym in a given grammar, n – the number of all possible arguments of the word v in this grammar
- l_i – the number of steps necessary to derive the word v (pseudonym) in the call wi
- k_i – the k-th step of the argument wi of the word v, $1 < k_i \leq l_i$
- R1, R2, …, R_j – successively ordered rules from the set of production of a given grammar, where Rj means the j-th rule from the ordered list of rules of grammar production, and the index j =1, 2, …, m positions the rule in a structured list of grammar production $\mathbf{G}_{\mathbf{B}}^{\mathbf{P}}$
- $\boldsymbol{R_j^{k_i}}$ – The j-th production rule used in step ki
- P (R_j) – probability of using the j-th rule from the production list

- $\boldsymbol{P_k^i}$ (R_j) – probability of using the production rule Rj in the i-th word v in step k
- P(H) – probability of compromising the pseudonym v as a result of an attack during the anonymisation process.

Let $\boldsymbol{\beta_{i^*}} = \mathbf{max}_{k_{i^*}} \boldsymbol{P_k^{i^*}}$ (Rj) for the given established i*, then we get the following inequality:

$$\begin{aligned}\mathrm{P(H)} &= \sum_{k_{i*}=1}^{l_{i*}} P_k^{i^*}(R_j^{k_{i*}}) \le \beta_{i^*} \cdot l_{i^*} \le \\ &\le \max_i \{\beta\mathrm{i} \cdot \mathrm{li}\} \le \max_i\{\beta_i\} \cdot \max_i \{l_i\} \le \\ &\le \max_j \mathrm{P(Rj)} \cdot \max_i \{l_i\} \le \mathrm{n} \cdot \max_j \mathrm{P(Rj)} \cdot \max_i \{l_i\} \le 1\end{aligned}$$

In the last inequality on the left side, there are two unknown parameter values n i $\mathbf{max}_i\{\boldsymbol{l_i}\}$. Even if the attacker had managed to gain knowledge about the definition of grammar – which would have been difficult due to its asymmetrical encryption system – the attacker would not have been able to estimate P(H) because of the two unknown parameters. However, in the case of explicit context-free grammars we have $\mathrm{n}=1$ and then we know the value of the parameter $\mathbf{max}_i\ \{\boldsymbol{l_i}\}$, because then $\mathbf{max}_i\ \{\boldsymbol{l_i}\} = \mathrm{l}_1$, there are no other trees of deductions, there is only one derivation of the word. If in such a situation an attacker would take over the definition of probabilistic unambiguous context-free grammar and would also have knowledge about its unique properties, they would encounter the problem of high computational complexity of the order $\boldsymbol{m^{l_i}}$, which even with the computational power of modern computers is a problem. It is not easy to find the argument of the word v and discredit the pseudonyms without having a sequence of digits representing the order of their production rules, because the person creating the pseudonyms can manipulate the values of parameters both m and li giving them appropriately large values. However, if the attacker was lucky and the blind hit enabled him to derive the word v, it would also know the parameter l, so they could easily estimate the probability P(H).

References

Act of 16 September 2011 on exchange of information with law enforcement agencies of the Member States of the European Union (In Polish: Ustawa z dnia 16 września 2011 r. o wymianie informacji z organami ścigania państw członkowskich Unii Europejskiej, państw trzecich, agencjami Unii Europejskiej oraz organizacjami międzynarodowymi, Dz.U. 2011/

230/1371). http://prawo.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20112301371 (accessed October 28, 2019).

Borucki, B. 2009. Methodology of privacy and security protection of the medical personal data (In Polish: Metodyka ochrony poufności i bezpieczeństwa medycznych danych osobowych). *Ultrasonografia* 36:9–20.

Danilowicz, C., and N. T. Nguyen. 2000. Consensus-based methods for restoring consistency of replicated data. In *Intelligent information systems. Advances in soft computing*, ed. M. Kłopotek, M. Michalewicz, and S. T. Wierzchoń, vol. 4, 325–36. Heidelberg: Physica-Verlag.

European Parliament and Council of the European Union. 2016. 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 (General Data Protection Regulation) (Text with EEA relevance). http://data.europa.eu/eli/reg/2016/679/oj/eng/.

FIPS 140-2. 1999. *Security requirements for cryptographic modules*. Gaithersburg, USA: NIST.

Grajek, M., and L. Gralewski. 2009. *Narodziny kryptologii matematycznej*. Warsaw, Poland: Semper.

Gruschka, N., V. Mavroeidis, K. Vishi, and M. Jensen. 2018. Privacy issues and data protection in big data: A case study analysis under GDPR. arXiv. IEEE, 5027-5033. https://arxiv.org/pdf/1811.08531.pdf.

Guidance Note: Guidance on Anonymisation and Pseudonymisation. 2019. Data Protection Commission. https://dataprotection.ie/sites/default/files/uploads/2019-06/190614%20Anonymisation%20and%20Pseudonymisation.pdf.

Guidelines on the use of cloud computing services by the European institutions and bodies. 2018. European Data Protection Supervisor. https://edps.europa.eu/sites/edp/files/publication/18-03-16_cloud_computing_guidelines_en.pdf.

Liber, A. 2014a. The issues connected with the anonymization of medical data. Part 1. The introduction to the anonymization of medical data. Ensuring the protection of sensitive information with the use of such methods as f(a) and f(a,b). (In Polish: Problemy anonimizacji dokumentów medycznych. Część 1. Wprowadzenie do anonimizacji danych medycznych. Zapewnienie ochrony danych wrażliwych metodami f(a) i f(a, b) anonimizacji). Medical Science Pulse 8 (1):13–21. doi:10.5604/01.3001.0003.3155.

Liber, A. 2014b. The issues connected with the anonymization of medical data. Part 2. Advanced anonymization and anonymization controlled by owner of protected sensitive data. (In Polish: Problemy anonimizacji dokumentów medycznych. Część 2. Anonimizacja zaawansowana oraz sterowana przez posiadacza danych wrażliwych). Medical Science Pulse 8 (2):13–24. doi:10.5604/01.3001.0003.3161.

Mostert, M., A. Bredenoord, M. Biesaart, and J. J. Delden. 2016. Big Data in medical research and EU data protection law: Challenges to the consent or anonymise approach. *European Journal of Human Genetics* 24 (7):956–60. https://www.nature.com/articles/ejhg2015239. doi:10.1038/ejhg.2015.239.

MSI-NET. 2016. Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications. Committee of Experts on Internet Intermediaries. https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a.

Nabywaniec, D. 2019. *Anonymization and masking of sensitive data in enterprises (In Polish: Anonimizacja i maskowanie danych wrażliwych w przedsiębiorstwach)*. Gliwice, Poland: HELION.

Opinion 05/2014 on Anonymisation Techniques. 2014. Data Protection Working Party, Article 29 of Directive 95/46/EC (0829/14/EN WP216). https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

Simon, S. 2001. *Księga szyfrów (The code book – The science of secrecy from ancient Egypt to quantum cryptography)*. Warsaw, Poland: Albatros.

Tripathy, B. K., Reddy, J. Manshua, G. V., and G. S. Mohisin. 2012. An efficient algorithm for anonymization of set-valued data and representation using FP-tree. *International Journal of Advanced Information Technology* 2 (5):1–14. doi:10.5121/ijait.2012.2501.

Wald, A. 1945. Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics 16 (2):117–86. doi:10.1214/aoms/1177731118.

Whitelegg, D. 2018. Minimizing application privacy risk. Practical application development techniques to alleviate risk. https://developer.ibm.com/articles/s-gdpr3.

Yu, S. 2016. Big privacy: Challenges and opportunities of privacy study in the age of big data. *IEEE Access* 4:2751–63. doi:10.1109/ACCESS.2016.2577036.